Investigation of New Strategies for Identifying Causal Mechanisms in Alzheimer's Disease Taking Bioinformatics Approaches Beyond GWAS

Emily Ann Baker

A thesis presented for the degree of
Doctor of Philosophy

School of Medicine,
Cardiff University,
United Kingdom,
September 2018

Declaration

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed ..... (candidate)

Date .....

Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed ..... (candidate)

Date .....

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated, and the thesis has not been edited by a third party beyond what is permitted by Cardiff Universitys Policy on the Use of Third Party Editors by Research Degree Students. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed ..... (candidate)

Date .....

Statement 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Acknowledgements

Firstly, I would like to acknowledge my supervisor Valentina Escott-Price. I was told before I started a PhD that a good supervisor is almost more important than the area of research you choose, and I am incredibly grateful to have had such a fantastic supervisor. She has been a huge support throughout the PhD and has given me the opportunity to develop over the four years. I particularly appreciate the glasses of wine, Russian jokes and games of Werewolf! Thank you!

I would also like to thank my other three PhD supervisors; Beccy Sims, Karl Michael Schmidt and Julie Williams. Thank you for all the support and proofreading! In addition I would like to acknowledge Peter Holmans and Michael O'Donovan, for proofreading many abstracts and helping to develop ideas presented in this thesis.

I would not have been able to pursue a PhD without generous funding from the MRC. The additional expenses budget has enabled me to attend a vast number of conferences which has been a great experience.

I would like to thank my parents, Jos and Brian, they have always been hugely supportive of me, both emotionally and financially, and I appreciate every sacrifice they both made for me to pursue my academic career. Also, thanks to my sister, Alice, and extended family and friends for always being able to take my mind off my research when I needed a break! You'll be pleased to know that after 10 years studying I am finally ready to give up being a student!

Finally to Nick, thank you for always believing I can achieve anything. Your unfaltering confidence in me never fails to encourage me; I would never have pursued, let alone completed a PhD if it wasn't for your constant support. But mostly, thank you for always managing to make me laugh when I needed it most! So, do you think I deserve a puppy now?

Abstract

The ideal outcome for medical research is the ability to provide personalised treatment for the population. Genetics studies are necessary to achieve this, since a person's genetic code does not change over time. Alzheimer's disease (AD) is a neurodegenerative disorder with a significant genetic component. Clinical trials in AD are difficult to design due to advanced neuropathological changes before development of symptoms, consequently, it would be beneficial to determine an individual's risk of AD.

Genome-wide association studies (GWAS) have unearthed over 20 variants associated with AD. It is expected there are more variants associated with disease which may explain disease aetiology, however, GWAS is not able to detect additional variants without increasing sample size. Set-based analysis is an attractive alternative to determining the association of one single nucleotide polymorphism (SNP), since the combined effect of all SNPs in the set may be captured. A gene-based analysis using the MAGMA approach in the AD data shows this by finding additional independent gene associations using identical data.

Polygenic Risk Scores (PRSs) are used for a variety of purposes in assessing the genetic liability to disorders. To further improve the power of set-based analyses, PRS as a set-based analysis is considered; this approach incorporates external data to improve power. This power increase is shown compared with other set-based methods using simulation studies and identifies two novel genes in imputed AD data; *CSMD1* and *MACROD2*.

The downside to the PRS approach is that it assumes independence between SNPs and thus data must be pruned for Linkage Disequilibrium (LD). Therefore, a novel approach which extends PRS and adjusts for LD is presented. This method is termed POLARIS and is shown to have better power compared to MAGMA in simulation studies and has determined three novel genes; *PPARGC1A*, *RORA* and *ZNF423*, in imputed AD data.

Nomenclature

p	P-value
Nsnps	Number of SNPs
p_{Fisher}	Fisher's set-based p-value
P	Probability
M	Number of SNPs
$\sum_{i=1}^{M}$	Sum of all values between 1 and M
$ln()$	Natural logarithm $log_e()$
χ^2_{2N}	Chi-squared distribution with 2N degrees of freedom
$Unif(0,1)$	Uniform distribution with mean 0 and variance 1
$p_{one-sided}$	One sided p-value
$p_{two-sided}$	Two sided p-value
p_{Simes}	Simes set-based p-value
$\min_{j=1,...,M}$	Minimum of all values between 1 and M
$p_{\_MAGMA}$	MAGMA's set-based p-value
β, BETA	Natural log of the SNP effect size ($log_e(OR)$)
g	SNP genotypes
π	Risk of disease
$p_{\_PRS}$	PRS set-based p-value
λ	Genomic control parameter
N_{genes}	Number of genes
N	Number of individuals
P_{sc}	Self-contained p-value
P_c	Competitive p-value
SE	Standard error
r^2	Measure of LD/correlation between 0 and 1
$N(a,b)$	Normal distribution with mean a and variance b
SD	Standard deviation
r	Correlation coefficient
$\tilde{g}$	SNP LD-adjusted dosages
C	Correlation matrix between SNPs
λ_k	Eigenvalues from spectral decomposition of the correlation matrix C
x_k	Orthonormal eigenvectors from spectral decomposition of the correlation matrix C
S	Covariance matrix between SNPs
x	Mean zero data
λ_0	Ridge parameter for POLARIS $\left(\sqrt{\frac{1}{N}}\right)$
I	Identity matrix
D	Correlation matrix between SNPs estimated from the test set
$\tilde{\beta}$	LD-adjusted βs
E[ ]	Expectation

Contents

List of Figures

List of Tables

1 Introduction

1.1 Introduction

The gold standard endpoint for medical research is to develop personalised treatment for all individuals; whether this is through the use of prevention strategies and lifestyle modifications or personalised medicines. Genetics studies are crucial to achieving this goal, with research extending from single gene disorders to complex disorders and genomics [1]. Major advances in genetic research have been influenced by the completion of the Human Genome Project [2], the development of Genome-Wide Association Studies (GWASs) [3] and the reduced cost of genome sequencing [4]; and the rate of discovery in this field is still accelerating [5].

1.2 Alzheimer's Disease

1.2.1 What is Alzheimer's Disease (AD)?

Late Onset (LO) Alzheimer's Disease (AD) is a devastating neurodegenerative condition with a significant genetic heritability [6]. AD is the most common form of dementia first described by Alois Alzheimer in 1907 [7]. LOAD is defined when the disease onset occurs after the age of 65. It is possible for the disease to develop at a younger age, this is termed Early Onset (EO) AD, and is a much rarer form of the disease [8].

AD is a growing public health concern, with the prevalence expected to increase in coming years. It currently affects more than 520,000 people in the UK [8] and 1 in 3 people in the

UK today will develop dementia in their lifetime. Globally there is expected to be a 204% increase in the number of people living with dementia by 2050 [9]. With growing disease prevalence comes growing economic impact; the cost in the UK is expected to more than double by 2040, and the impact of dementia on the economy is larger than the combined cost of cancer and heart disease [9]. In the US, 7.9 trillion dollars of medical costs could be saved by early diagnosis [10].

1.2.2 Symptoms

AD begins as memory loss and develops into severe dementia and death. The disease may present differently in each person but can encompass a number of varied symptoms. The disease may progress differently as well, and the life expectancy after initial symptoms is on average 8-10 years, although again this varies a lot, particularly with the age of onset [8].

Usually, memory loss begins as memory lapses where people with AD struggle to learn new information or remember recent events [8]. Sometimes memory loss which is due to dementia can be confused with forgetfulness due to normal aging [9]. This memory loss impacts daily life because people with AD may forget medication, appointments, conversations or events and can forget their local area and become confused with directions. In early stages, Alzheimer's patients tend to have less trouble remembering events which happened a long time in the past. This memory loss is caused by damage to the hippocampus [8].

Alzheimer's patients may also experience difficulties with thinking, communication, reasoning and perception [8]. For example, they can have issues with problem solving, judging distances, changes in sleep pattern or may have mood changes such as becoming irritable and anxious [8][9].

The symptoms become more severe as the disease progresses, resulting in the person with AD requiring much more support and care and eventually leading to them needing help with all daily activities.

1.2.3 Diagnosis

There is no easy test to determine whether a person has AD, particularly because the symptoms are common to a number of other diseases. AD is predominantly assessed by interviewing the person, and if possible a close relative or friend. The symptom development is thoroughly investigated to determine if memory has got worse gradually over time. Pen-and-paper tests are used to assess the individuals mental ability and sometimes a brain scan is used, usually in order to exclude other conditions [8].

Overall, these different assessments made by GPs, psychiatrists or neurologists can be collated to make a decision about whether the person has AD or not [8][9]. Although it is not possible to be certain about a diagnosis without determining if the neuropathology of AD is present by autopsy [11].

1.2.4 Neuropathology of Alzheimer's Disease

Pathologically, AD is distinct and recognizable [11]. AD is defined by the presence of intracellular neurofibrillary tangles and extracellular amyloid plaques [12]. The neurofibrillary tangles are made up of an unusual build up of phosphorylated tau in the neurons. The amyloid plaques have a beta-amyloid core encased by neurites [11].

One of the major challenges in identifying treatments for AD is the amount of pathological change in the brain prior to an individual developing symptoms, this is shown in Figure 1.1. Therefore, by the time AD is diagnosable, substantial neurodegeneration has already occurred. Additionally, when determining control subjects for studies (those without AD), it is possible that these subjects may go on to later develop AD, or may already have some of the pathological changes which characterise the disease- thus reducing the power of these studies. Genetics is therefore hugely beneficial in this case, since a person's genetics could be assessed at birth. If it is possible to accurately predict individuals who are most likely to develop AD, treatments can be targeted at these individuals prior to symptom development.

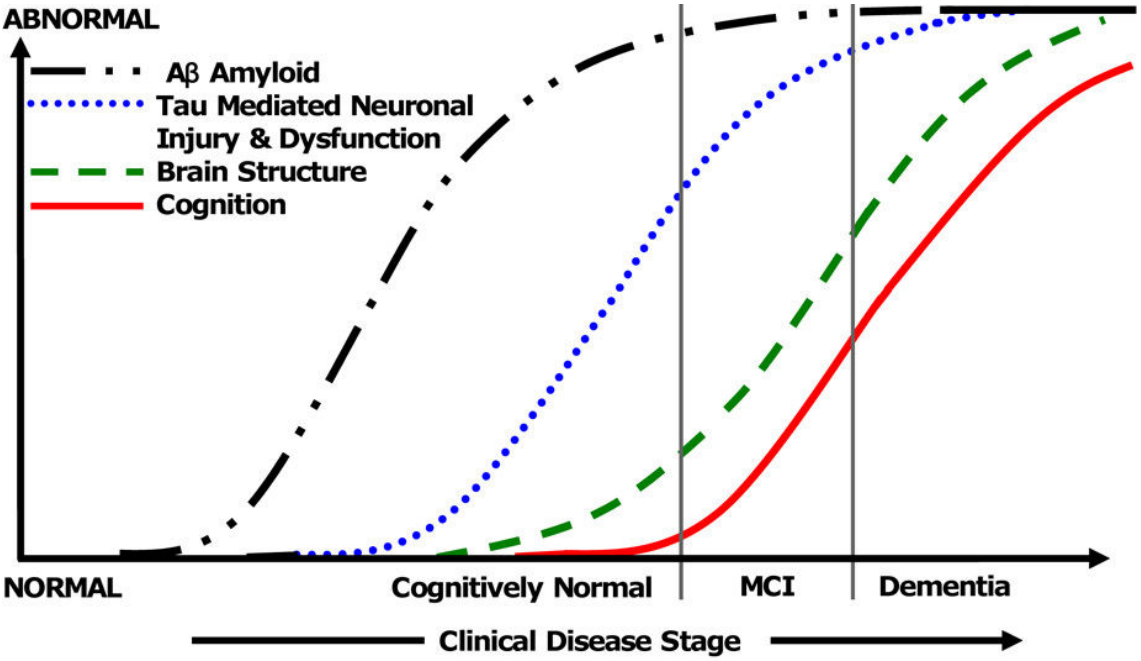


Figure 1.1: AD Pathology Progression [13]

1.2.5 Treatments

The cause of sporadic LOAD is unknown and as such there are no treatments specifically for AD, but current treatments are targeted at specific AD symptoms or to slow disease progression [14]. The benefits of medicines for AD are small, and a number of non-drug therapies can also be used to help care for a person with AD [8].

There are two main drug treatments used in AD; these are Acetylcholinesterase inhibitors and NMDA receptor antagonists. AD patients have reduced levels of acetylcholine which passes messages between nerve cells. Acetylcholinesterase inhibitors prevent an enzyme from breaking down acetylcholine in the brain. NMDA receptor antagonists stop the effects of glutamate. Glutamate also passes messages between nerve cells, this is released in excess in AD brains leading to additional brain damage [8].

Non-drug treatments such as maintaining activity and brain stimulation in the person with AD can be beneficial. This involves finding activities suited to each person's abilities, including arts, crafts, physical exercise and music. In addition, maintaining good social interactions can help to reduce symptoms [8].

1.2.6 Genetic Risk Factors

Firstly, the genes related to EOAD were determined, using gene-mapping [12]. These genes were *APP* [15], *PSEN1* [16] and *PSEN2* [17].

However, the most common form of AD is LOAD which is often defined as having an onset after the age of 65, LOAD is also the primary focus of this thesis. LOAD is a complex genetic disorder, meaning that the disease does not follow Mendelian inheritance. The apolipoprotein E (*APOE*) gene is the strongest genetic risk factor for LOAD [18]. *APOE* has three different isoforms: $\epsilon 4$, $\epsilon 3$ and $\epsilon 2$. The number of $\epsilon 4$ isoforms is much higher in cases compared to controls [12].

GWAS, gene-based analyses and whole-exome sequencing studies have identified a large number of genes which are associated with LOAD, these are discussed in more detail in Section 1.3. There are over 20 genes which have been found to have an association with AD [19][20][21][22]. In addition, eight biological pathways have been found to be associated with AD, including immune response and cholesterol transport [23][24].

1.2.7 Environmental Risk Factors

This thesis focuses on the genetic risk in AD, but a number of environmental risk factors have also been found to be associated with AD [25].

The largest risk factor for AD is age; it mainly affects people over the age of 65. Over 65, the risk of AD doubles every 5 years. It is found that AD is more common in females compared to males, although the reason for this difference is unclear, but is hypothesised to be related to the lack of oestrogen after menopause [8].

Risk factors for dementia are also risk factors for cardiovascular disease, so maintaining a healthy lifestyle and exercising regularly may reduce the risk of dementia. A person can maintain a healthy lifestyle by not smoking, taking regular exercise, eating healthily and maintaining healthy blood pressure [9].

Additional environmental factors strongly associated with AD are air pollution, for example, increases in nitrogen oxide, particulate matter and ozone levels increase the risk of AD, being in the presence of pesticides or solvents at work, and increased levels of vitamin D also lead to the increase in AD risk [25].

1.3 Methodological Approaches to Genetic Data

Methodologies in genetic research have advanced in recent years. This has been aided by improved computational resources, advanced laboratory methods and increased sample sizes in genetics studies.

These methods are discussed in AD and also Type II Diabetes and Schizophrenia (SZ). The findings in these complex traits are also discussed since Type II Diabetes is a risk factor for AD [26] and psychosis is presented in some people with AD [27]. Genetic results from these two disorders may aid in the biological understanding of AD.

1.3.1 Genome-Wide Association Study (GWAS)

Genetic discoveries reached a plateau due to studies being underpowered for the effect size of each individual genetic variant [3]. Therefore, GWASs which assess whether genome-wide variants are associated with disease in a number of individuals, were developed to find associations between complex traits and genomic loci; specifically to determine associations between common diseases and Single Nucleotide Polymorphisms (SNPs) [3]. The aim was to use detected variants to better understand the biology of disease, however, translating GWAS results to informing disease biology is not straightforward [28]. In order to have power to detect such small associations with disease, a large number of subjects were required and a large number of SNPs across the genome would be analysed.

The Genetic and Environmental Risk in Alzheimer's Disease Consortium (GERAD) published a Genome-Wide Association Study (GWAS) that identified novel variants in *CLU* and *PICALM* which were associated with AD [19], see Figure 1.2. Concurrently, the

The European Alzheimer's Disease Initiative (EADI) identified the *CR1* and *CLU* loci to associate with AD [20]. Subsequent publications by Genetic and Environmental Risk in Alzheimer's Disease Consortium (GERAD), the Alzheimer Disease Genetics Consortium (ADGC) and The Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium (CHARGE) Consortium identified a further 5 novel loci [29][30][31].

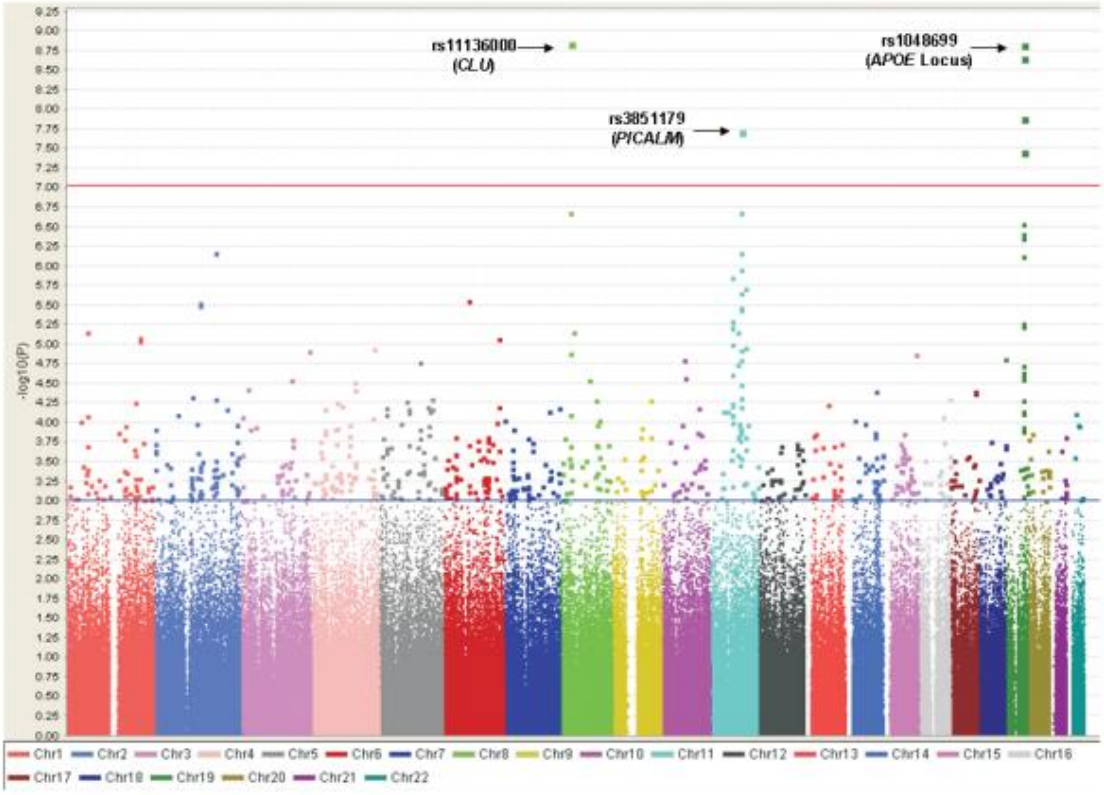


Figure 1.2: Manhattan Plot From Harold et al. 2009 [19]

Due to the large sample sizes required to detect such small individual SNP effects, a number of different consortia collaborated to form one large consortium for many different complex traits. This involved the meta-analysis of SNP effects from each group in order to increase power and detect a larger number of variants.

The International Genomics of Alzheimer's Project (IGAP) [20] consortium is an amalgamation of the four different genetic groups (GERAD, EADI, ADGC and CHARGE) previously discussed. Meta-analysis of the 4 GWAS datasets determined 11 novel variants associated with AD, these results are shown in the Manhattan plot in Figure 1.3 [20].

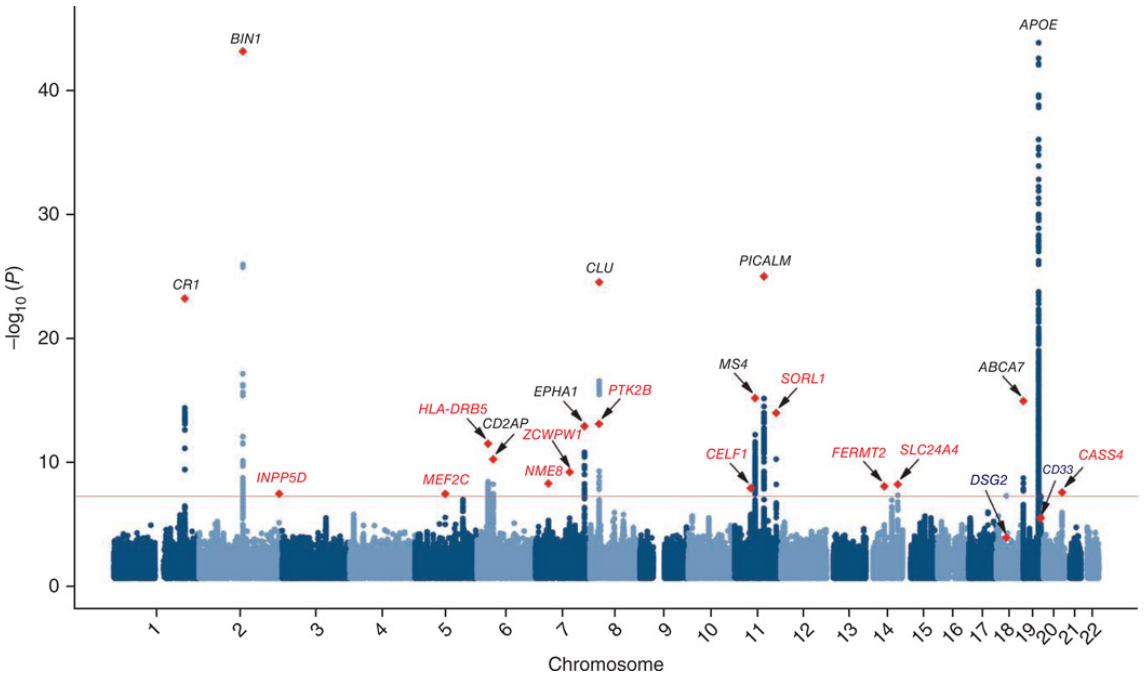


Figure 1.3: Manhattan Plot From Lambert et al. 2013 [20]

SZ is another disorder where a large number of variants have been identified using GWAS and this has led to a much greater understanding of disease aetiology. A GWAS of 3,416 subjects (479 cases, 2,937 controls) and a replication set of 16,726 subjects determined a strong independent association with *ZNF804A*, which has a potential role regulating gene expression [32]. Five novel loci associated with SZ were determined using a GWAS with a much larger number of 21,856 subjects and an additional independent replication sample with 29,839 subjects, where subjects were of European ancestry. The most associated loci was a neuronal development regulatory gene; *MIR137* [33]. The identification of an additional 13 loci was possible using a multi-stage GWAS where a Swedish sample was used and meta-analysed with another SZ GWAS [34]. Finally, an extremely large GWAS including 36,989 SZ cases and 113,075 controls found 108 novel loci, and thus demonstrated the ability for large GWAS to detect a huge number of associations [35].

GWAS has also been particularly successful in studies of Type II Diabetes [28]. The use of GWAS has enabled the identification of 38 SNPs associated with Type II Diabetes; these associations were determined against a dichotomous variable, but when considering continuous glycemic traits, an additional 24 variants were discovered [36].

Initially, researchers were sceptical about the validity of the experimental design of GWAS, but it is much more widely accepted more recently due to the number of robust findings [28].

The GWAS design has led to the discovery of a huge number of associated variants across a range of disorders, GWAS are not hypothesis driven, all genotyped SNPs are tested for their association with a particular disease. However, there are some potential issues with the GWAS approach. The first being that GWAS studies have been unable to explain the total genetic variation estimated by linkage studies. This 'missing heritability' could be due to interactions between SNPs, rare variants or variants which have not been genotyped or are imputed with poor accuracy. Additionally, annotations of identified variants are characterised using the most proximal gene, but the variant may be acting upon an entirely different gene [36]. Another issue with GWAS is whether the associated variants are biologically relevant, although the number of biological pathways implicated in diseases, based on GWAS results, is substantial [3]. Linking GWAS results to the underlying disease biology has been successful by using new analytical techniques, molecular technology and additional data [28].

Meta-analysing the results of different GWAS is very beneficial in terms of increasing the power to determine SNPs associated with disease, although, issues may arise where the GWASs have heterogeneity between populations; differing Linkage Disequilibrium (LD) structure between SNPs may lead to different SNPs being statistically significant in each GWAS. Meta-analysing these SNPs together may cancel their individual effects into a null overall effect.

1.3.2 Polygenic Risk Score (PRS)

GWAS unearthed a large number of genetic associations with complex disorders, all of which have a small individual effect. However, there were still potentially vast numbers of variants which did not reach genome-wide significance, but may still collectively contribute to disease risk.

In order to determine the polygenic effect of a particular disease, Polygenic Risk Score (PRS) was developed. Polygenic Risk Score (PRS) involves the combination of all SNPs across the genome into one risk score for each individual. It is then possible to assess the genetic burden of a particular disorder and determine whether the risk score is able to predict whether an individual will have the disease [37].

The effectiveness of PRS was first shown in SZ, where the polygenic component was highly associated with SZ ($p = 1.9 \times 10^{-19}$) and explains at least one-third of liability variation [37]. PRS was then calculated using the largest Schizophrenia (SZ) GWAS and was able to confirm that PRS is associated with SZ and is able to predict case/control status. In this data, 7% of variation on the liability scale is explained using PRS [35].

AD was also shown to be a complex polygenic disorder, with the polygenic component strongly associated with AD ($p = 4.9 \times 10^{-26}$) [38]. Additionally, the best reported accuracy to predict whether or not a subject has AD based on PRSs is 78.2%. This accounts for 90% of the maximum prediction accuracy possible for AD [39].

PRS computed in 18 genetic loci associated with Type II Diabetes were successfully able to predict diabetes cases, but this provided only a slight improvement compared to using known risk factors only [40]. However, this prediction was further improved upon by the inclusion of a large number of SNPs (Nsnps=1,000), this was found in the Estonian Biobank cohort [41].

PRS has the main advantage that it is able to produce a risk score per subject, which can be used in further analyses. PRS is also able to increase power from a standard GWAS by incorporating the SNP effect sizes from an independent study. It improves power by accounting for a number of SNPs which individually have a small effect size, but collectively contribute to disease risk. In addition, it is desirable to predict a person's risk of disease using PRS, as this can be used in clinical trials or to prioritise subjects to follow up. It may also enable subjects to be treated for a particular disease in a precautionary manner before they are actually diagnosed with the disease (particularly relevant to adult-onset disorders). Disease prediction has the potential to be used as a tool for personal health management [42].

PRS assumes independence between SNPs and therefore, data must be pruned for Linkage Disequilibrium (LD) before analysis. LD is a correlation between SNPs, or the non-random association of alleles at different loci [3]. This likely removes a large number of SNPs from the data, and some (albeit small) LD between SNPs may remain. PRS requires the use of two independent datasets, one of which must have individual genotype data. Unfortunately, individual genotype data is rarely freely available to download, so this may limit researchers able to use this approach. The use of PRS in precision medicine has a number a challenges; firstly, the lack of genetic research in non-European populations mean that results are not widely applicable, which is a necessity for clinical applications, and secondly, education for the public and clinicians to use the new concept of a continuous genetic risk component, rather than the presence/absense of a particular variant [43].

1.3.3 Gene-Based Analysis

Gene-based analysis offers an attractive alternative to single SNP analyses, since the combined effect of SNPs within the gene may be captured, whereas single SNP analyses are often underpowered due to the small effect sizes of individual SNPs. In addition, gene-based analysis specifically identifies the gene associated with disease rather than a single SNP as a proxy for the gene.

Genes are found to be fairly consistently associated with disease across different populations using a gene-based analysis. In contrast, different SNPs in a set in LD may be found to be associated with a disease in different samples. Gene-based analyses also directly provide information for functional analysis [44].

There are a number of methods to assess the gene-based effect by combining the effects of all SNPs within the gene. Fisher's method [45] combines p-values of all SNPs in the gene to generate an overall gene p-value, but assumes independence between SNPs. Simes [46] is similar to Fisher's method, except it adjusts the SNP p-values for the number of SNPs in the gene and tests whether at least one of these SNPs is associated with disease. GATES extended Simes [44] is an addition to Simes method which also incorporates functional information. Rank/threshold truncated products of p methods [47][48][49] forms the product

of the most significant SNPs in the gene. Set-based analysis as implemented in PLINK [50][51] where a permutation procedure is used to adjust SNPs for LD. Brown's method [52][53] is an extension to Fisher's method but which adjusts for the LD structure between SNPs. A logistic kernel-machine based test is available which accounts for epistatic and non-linear SNP effects for all SNPs within the gene [54]. Multi-marker Analysis of GenoMic Annotation (MAGMA) [55] uses a regression based approach which computes principal components from all SNPs to adjust for LD. Finally, Pascal [56] utilises the sum and maximum of chi-squared statistics to generate a gene score.

A gene-based analysis has been undertaken in the International Genomics of Alzheimer's Project (IGAP) AD data using Brown's method [52]. This approach determined two additional novel genes; *TP53INP1* and *IGHV1-67* [21]. These results are shown in the Manhattan plot in Figure 1.4.

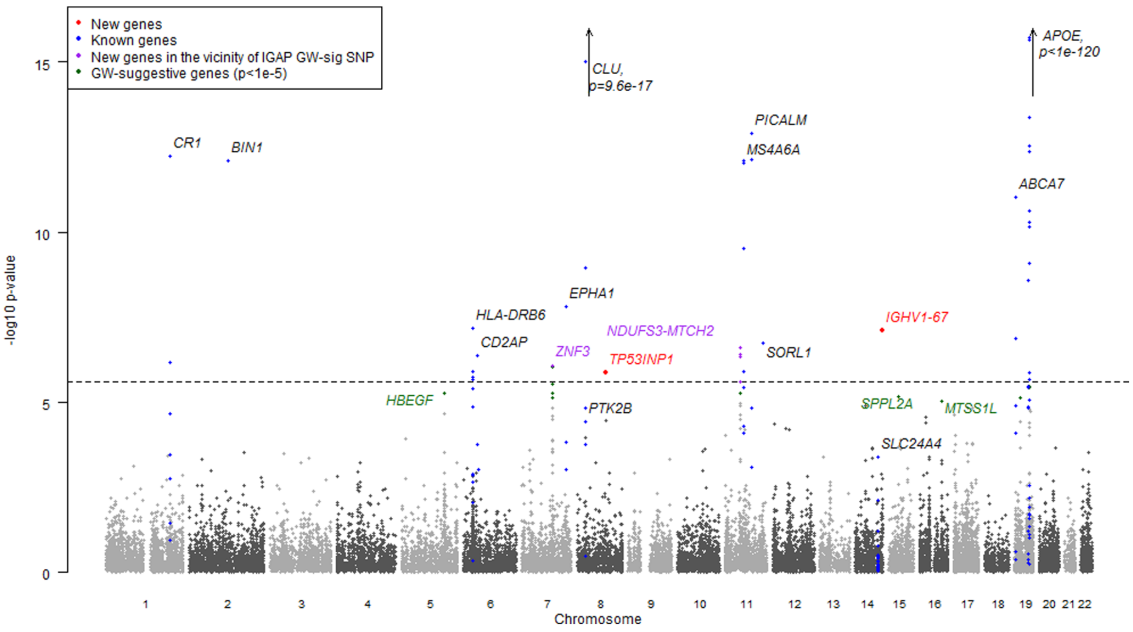


Figure 1.4: Manhattan Plot From Escott-Price et al. 2014 [21]

Gene-based analyses in Frontotemporal Dementia (FTD) have aided in identifying gene associations which were undetermined using GWAS only, thus demonstrating the increased power of gene-based approaches. In 3,348 FTD cases and 9,390 controls, *APOE* and *TOMM40* were associated with behavioural variant FTD and *SERPINA1* and *ARHGAP35* were found to be associated with progressive non-fluent aphasia [57].

The power of gene-based analyses is further demonstrated in Rheumatoid Arthritis (RA), where a gene-based analysis in publically available RA datasets (14,361 cases, 43,923 controls of European ancestry and 4,873 cases, 17,642 controls of Asian ancestry) determined 221 novel genes associated with RA, 71 of which overlapped both ancestries [58].

Gene-based analyses combine the information of all SNPs in a gene, so are more likely to find novel associations and more informative results [59]. Genes are more robustly determined across different populations compared to single SNP analysis, since due to LD, associated SNPs may be representing different causal SNPs [44]. Restricting the analysis to genes makes biological sense, since a gene is a defined functional unit of the genome [44]. Gene-based analyses require a less stringent multiple testing corrected significance threshold compared to single SNP analyses, due to fewer genes being assessed compared to SNPs.

There are a huge number of potential gene-based approaches, so there may be difficulty in choosing the appropriate method based on the data available to the researcher. It could also be difficult to annotate SNPs to genes, for example, should only SNPs within the gene be included, or is it more appropriate to include a window of SNPs around the gene which may contain additional transcriptional regulatory elements.

1.3.4 Pathway Analysis

Despite the findings of a large number of SNPs and genes which are associated with disease, the biology underlying the aetiology of disease remains difficult to determine. Therefore, a large number of methods have become available in order to link genetic associations to a relevant biological process, these are known as gene-set or pathway analyses [60].

Pathway analysis is another set-based approach where a group of genes which all have a similar biological function are used, it is then investigated whether this group of genes is associated with disease. A pathway analysis is beneficial because it gives specific biological information about the potential mechanisms involved in disease.

A pathway analysis has been undertaken in IGAP AD data. This used the Association

LIst Go AnnoTatOR (ALIGATOR) algorithm [61] which defines genes to be significant if they contain a single SNP with a p-value less than a set threshold. The significant gene set is then compared to randomly generated gene sets to determine if there is an excess of association with AD in the gene set. Eight pathways were found to be associated with AD [23][24]; these pathways are immune response, regulation of endocytosis, cholesterol transport, hematopoietic cell lineage, proteasome-ubiquitin activity, reactome hemostasis, clathrin and protein folding.

Five pathways have been found to be associated with schizophrenia; serotonergic synpase, ubiquitin mediated proteolysis, hedgehog signalling, adipocytokine signalling and renin secretion. These pathways were consistent across three different populations of European-American, African-American and Han Chinese ancestry [62].

In addition, the presense of pathways associated with Type II Diabetes have also been determined. These pathways are tight junction, cell cycle, melanogenesis, vibrio cholerae infection, gastric acid secretion, phagosome, ubiquitin mediated proteolysis and protein processing in endoplasmic reticulum [63].

Pathway analyses allow the connection of genetic associations with their relevant biological processes in order to better understand disease mechanisms. Pathway analysis can, like gene-based analyses, further improve power by assessing the collective effect of SNPs or genes with small individual effects. Pathway analyses may also account for genetic heterogeneity within populations [60].

There are a large number of pathway approaches available and no particular consensus within the field regarding the optimal approach to take. Pathway analyses are hugely dependent on these methods, and how the gene-sets are defined, and thus results should be interpreted with care [60].

1.3.5 Rare Variant Analysis

It was thought that one reason for the 'missing heritability' from GWAS could be due to the presense of rare variants which are associated with disease, but are too rare to be

determined from a GWAS.

In AD, low frequency risk variants have been identified through next generation sequencing (*TREM2*) [64] and an whole-exome association study (*PLCG2*, *TREM2* and *ABI3* [22]). These three genes are highly expressed in microglia, which further implicates the immune response pathway, which was identified from pathway analyses.

Whole exome sequencing studies have determined novel genes associated with a number of different disorders. *PLA2G4E* has been identified as a potential gene associated with panic disorder in a Japanese population, although it did not reach gene-wide significance [65]. In SZ cases, rare variants found from a whole exome sequencing study show enrichment in loss of function intolerant genes ($p < 3.6 \times 10^{-10}$) [66].

The study of rare variants is advantageous since it is possible to explain additional disease risk and heritability. The importance of rare variants in disease is well known, with highly penetrant rare variants causing rare subtypes of common diseases. Advances in genetic technologies has allowed the successful sequencing of low frequency variants for a reasonable cost [67].

Cost is, however, still somewhat prohibitive, and rare variant studies tend to focus on exome regions, although whole genome sequencing is likely to become more widely used as costs continue to decrease [67]. It will be a challenge to develop appropriate statistical methods which can differentiate functional variants from rare neutral variants [68].

1.4 Aims

The main aim of this thesis is to identify novel genes which are associated with AD in order to better understand the aetiology of AD. This will be done by researching recent, powerful methods which are better able to identify novel variants than GWAS alone. If current methods are limited, then the development of a novel approach which improves power will be investigated. The implication and credibility of any novel variants found to be associated with AD will be assessed and any potential biological mechanisms which

may explain the development of AD will be discussed.

1.5 Thesis Outline

The second chapter in this thesis discusses the two main datasets used throughout the thesis and the different versions of this data. In addition, the main methodological approaches which are predominantly used in the thesis are outlined and the software which is used for the analyses is highlighted.

Next, a gene-based analysis in the AD data is carried out, using a widely used methodology, MAGMA, which has greater power than other gene-based approaches. This is to determine whether gene-based approaches have more power compared to single-SNP analyses and are thus able to find novel associations in the same data.

A new approach to gene-based analysis was then investigated which incorporates additional data to further improve power using the Polygenic Risk Score (PRS) methodology. This was tested against current methods; MAGMA, Fisher's and Simes. The PRS gene-based approach was then used in AD data to determine whether any novel genes are identified.

This PRS gene-based method is extended to adjust for correlation or LD between SNPs, this novel method is called POlygenic Ld-Adjusted Risk Score (POLARIS). Again, this novel method was compared to MAGMA in simulated data and then applied to real AD data to identify potential novel genes.

POLARIS was then extended to compute a risk score across the whole genome, more similarly to standard PRS and was compared to another LD adjustment approach, LDpred. The ability of the POLARIS score to predict AD case/control status was assessed and compared to the prediction ability of PRS.

All risk score methods so far have been weighted using effect sizes from the same disorder; the final chapter investigates the use of weights from different disorders in a POLARIS gene-based analysis. This will enable the determination of genes in common between AD and other disorders, which may lead to a better understanding of the underlying biological

mechanism of AD.

2 Methods

2.1 Data

2.1.1 Genetic and Environmental Risk for Alzheimers Disease (GERAD) Data

The AD data used in this thesis was obtained from the Genetic and Environmental Risk in Alzheimer's Disease Consortium (GERAD) [19]. The GERAD stage 1 sample comprised up to 3,941 AD cases and 7,848 controls. These samples were recruited by the Medical Research Council (MRC) Genetic Resource for AD (Cardiff University; Kings College London; Cambridge University; Trinity College Dublin), the Alzheimers Research UK (ARUK) Collaboration (University of Nottingham; University of Manchester; University of Southampton; University of Bristol; Queens University Belfast; the Oxford Project to Investigate Memory and Ageing (OPTIMA), Oxford University); Washington University, St Louis, United States; MRC PRION Unit, University College London; London and the South East Region AD project (LASER-AD), University College London; Competence Network of Dementia (CND) and Department of Psychiatry, University of Bonn, Germany and the National Institute of Mental Health (NIMH) AD Genetics Initiative. All AD cases met criteria for either probable (NINCDS-ADRDA, DSM-IV) or definite (CERAD) AD. All elderly controls were screened for dementia using the MMSE or ADAS-cog, were determined to be free from dementia at neuropathological examination or had a Braak score of 2.5 or lower. For this GERAD data, the full individual genotypes for all individuals are available. Publically available control data from the Wellcome Trust Case Control Consortium (WTCCC) were also included in the data. After Quality Check (QC), the GERAD data used in this thesis has 13,164 subjects; 3,332 AD cases and 9,832 controls.

This GERAD data contains a total of 419,048 genotyped SNPs. This is termed the GERAD genotype data throughout this thesis.

The Haplotype Reference Consortium (HRC), version r1.1 2016, was used to impute GERAD genotype data on the Michigan Imputation Server [69], which to date, allows the most accurate imputation of genetic variants. Imputed genotype probabilities (also known as dosages) were converted to the most probable genotype with a probability threshold of 0.9 or greater. SNPs were removed if: their imputation INFO-score$<$ 0.4, minor allele frequency (Minor Allele Frequency (MAF))$<$ 0.01, missingness of genotypes$\geq$ 0.05 or Hardy Weinberg Equilibrium (HWE)$< 10^{-6}$. A total of 6,119,694 variants were retained. To correct for population structure and genotyping differences, all analyses were adjusted for gender and the top 15 principal components. This data is hereafter referred to as the GERAD imputed data.

2.1.2 International Genomics of Alzheimer's Project (IGAP) Data

Additionally, summary statistics from the International Genomics of Alzheimer's Project (IGAP) study [20] were used. IGAP is a large two-stage study based upon GWAS on individuals of European ancestry. In stage 1, IGAP used genotyped and imputed data on 7,055,881 SNPs to meta-analyse four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The Genetic and Environmental Risk in AD consortium - GERAD, The European Alzheimer's disease Initiative - EADI, the Alzheimer Disease Genetics Consortium - ADGC and The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium - CHARGE). In stage 2, 11,632 SNPs were genotyped and tested for association in an independent set of 8,572 AD cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1 and 2.

2.2 Methodological Approaches

2.2.1 Fisher's method

Fisher's method [45][53] is a method of combining the p-values of SNPs within a set to determine the evidence against a common null hypothesis [70]. It obtains an overall p-value for each set from a group of p-values which are independent tests of the same hypothesis. Fisher's set-based p-value is calculated using Equation 2.1.

$$p\_Fisher = P\left(-2\sum_{i=1}^{M} \ln(p_i) \geq \chi^2_{2M} \right) \tag{2.1}$$

where p_Fisher is Fisher's set-based p-value, P is the probability, M is the number of SNPs in the set and p_i is the p-value for SNP i.

If the summary statistics used to calculate the Fisher's statistic have been calculated adjusting for population covariates, then the overall set p-value will also be adjusted for population covariates.

Fisher's method assumes independence between SNPs, and therefore, in the presense of LD between SNPs, it is expected that the set-based p-value produced by Fisher will be biased since the type I error rate is likely inflated [44]. Fisher's method also assumes that under the null hypothesis the p-values follow a Uniform distribution, $p_i \sim Unif(0, 1)$. Provided that these assumptions hold, Fisher's method is asymptotically optimal.

Fisher's method has one disadvantage in that it treats large and small p-values asymmetrically, and therefore the method is asymmetrically sensitive to small p-values when compared to large p-values, for example if you combine two studies, one with p=0.001 and one with p=0.999, the combined Fisher's p-value would be p=0.008, therefore, the small p-values have a greater influence on the combined p-value. This asymmetry can result in bias when combining studies on the same null hypothesis [70].

2.2.2 Simes method

Simes' [46], like Fisher's, is a set-based method accounting for the number of SNPs within the set, however, it also attempts to correct for LD between SNPs in the set. The null hypothesis of this method is that no SNP within the set is associated with disease, and the alternative hypothesis is that at least one SNP within the set is associated with disease [44]. The Simes' set-based p-value is calculated using Equation 2.2. Simes removes the issue of LD by considering whether at least one SNP in the set is associated with disease, rather than combining the effect of all SNPs.

$$p\_Simes = \min_{j=1,...,M} \left(\frac{M p_{(j)}}{j} \right) \tag{2.2}$$

where p_Simes is the Simes' set-based p-value, M is the number of SNPs in the set, $p_{(1)}, ..., p_{(M)}$ are the SNP p-values arranged in ascending order, j is the rank and $p_{(j)}$ is the j^{th} p-value.

As with Fisher's method, p_Simes will be adjusted for population covariates if the summary statistics have been calculated with population covariate adjustment.

In essence, the Simes method corrects for multiple testing in such a way that is less conservative than the Bonferroni correction [71]. Like Fisher's method, Simes assumes that, under the null hypothesis, the set of p-values comes from a Uniform distribution, $p_i \sim Unif(0, 1)$ [46]. If the SNPs are independent, then the Simes set-based p-values are expected to also follow a Uniform distribution between 0 and 1, however, if SNPs are positively correlated, the estimate of the set-based p-value is likely to be conservative [44].

2.2.3 MAGMA

MAGMA [55] is a regression based approach which fully accounts for LD between SNPs. The matrix of SNPs within the set is transposed into Principal Components (PCs), and PCs with small eigenvalues are removed. The remaining PCs are then regressed against the phenotype of interest and an F-test is used to determine the strength of the associ-

ation between the set and the phenotype and thus gives the MAGMA set-based p-value, p_MAGMA. MAGMA also has the functionality to calculate set-based p-values using summary statistics (MAGMA-SUMMARY).

2.2.4 Polygenic Risk Score (PRS)

For M SNPs in a set, PRS [37] combines single-SNP genotypes g_i $(i = 1, \ldots, M)$ into a single regression predictor using single-SNP effect sizes ($\log(\text{Odds Ratio (OR)}_i) = \beta_i$) taken from a previous study as coefficients,

$$PRS = \sum_{i=1}^{M} \beta_i g_i = \beta^T g. \tag{2.3}$$

The PRS method implements a 2-stage approach, where independent discovery and test sets are available. The effect sizes β are determined from the discovery set and a vector of the number of risk alleles g is obtained from the test set. The underlying assumption is that individual genotypes are available for the test set, but only summary data (effect sizes β) for the discovery set are available.

The PRS method assumes that SNPs are independent and does not adjust for LD between markers and thus requires LD pruning [72].

Once the PRSs have been generated, a p-value is calculated using logistic regression, adjusting for population covariates. The logistic regression model can be seen in Equation 2.4.

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 PRS + \text{population covariates} \tag{2.4}$$

where π is the risk of disease, β_0 is the intercept and β_1 is the effect size for the association of the PRS with disease, often interpreted as the log(OR).

The strength of association between the PRS and disease is given by p_PRS, which is the p-value for the β_1 coefficient in the logistic regression model.

2.3 Software

2.3.1 R

R [73] is a free software which is widely used for statistical computing and graphics. R is the software used for all analyses in this thesis, unless stated otherwise.

2.3.2 PLINK

PLINK [50][51] is an open source whole genome association toolset which is used for a number of different genotype analyses throughout this thesis. In all cases, PLINK v1.9 was used.

2.3.3 Python

Python [74] is another programming language which is freely available. This is not specifically aimed at statistical programming, but it does have this functionality. It was used when parallel processing was required, since this language is easier to program parallel jobs compared to R. The Anaconda distribution (`https://www.anaconda.com`) of Python was used, since this optimises the processing of large datasets; this is a necessary requirement in genetics studies.

2.3.4 MAGMA

The MAGMA [55] approach comes with a corresponding software written in C. This is operated from the command line and enables the user to use either the MAGMA-Principal Component Analysis (PCA) or MAGMA-SUMMARY approach, depending on the data. MAGMA v1.06 was used throughout this thesis.

3 MAGMA Gene-Based Analysis in AD Data

3.1 Introduction

Set-based analysis offers an attractive alternative to single SNP analyses, since the combined effect of SNPs within the set may be captured. Single SNP analyses are often underpowered due to the small effect sizes of individual SNPs, set-based analysis considers the combined effect of all SNPs within the set, which may have a larger combined effect size and hence higher power to detect association than any individual SNP. In addition, gene-based analysis, a gene centred equivalent of set-based analysis, identifies genes associated with disease rather than a single SNP as a proxy for the gene. In gene-based analyses, genes are found to be fairly consistently associated with disease across different populations. In contrast, different SNPs in a set in LD may be found to be associated with a disease in different samples. Gene-based analyses also directly provide information for functional analysis [44]. Set-based analysis can also be employed as a pathway analysis, and applied to sets of SNPs defined by epigenomics for different tissue/cell types.

MAGMA v1.06 [55] is a recent approach which has emerged as a widely used and computationally efficient set-based method. It is a regression based approach which accounts for LD between SNPs when individual genotype data are available. The matrix of SNPs within the set is decomposed into PCs, and PCs with small eigenvalues are removed. The remaining PCs are then used as uncorrelated predictors in regression against the phenotype of interest and an F-test is used to determine the strength of the association between the set and the phenotype, providing the MAGMA gene-based p-value. The MAGMA program can be used on individual genotypes using the PCA method and also on sum-

mary statistics using Brown's method [52]. It often happens that summary statistics are available from a large consortium, while in-house studies with individual genotypes have smaller sample sizes. In such situations, the options for using MAGMA are either applying the PCA method to the in-house genotype dataset, or applying the summary statistic approach to a meta-analysis of summary statistics of both datasets. It has been shown as being more powerful than other gene-based softwares [55] such as PLINK [50][51] and VEGAS [75].

3.1.1 Objectives

The aims of this chapter are to:

- Investigate whether gene-based analysis has more power than single SNP analyses by combining the effect of all SNPs in the gene. MAGMA software is used to carry out these gene-based analyses.

- Determine whether it is possible to find equivalent signals using a gene-based analysis in smaller data compared to a single SNP analysis in larger data.

- Consider whether we can enhance our analysis using flanking regions around the gene, since transcriptional regulatory elements are likely to be contained within these intervals and there may be some merit in capturing the variation within these regions [76].

- Demonstrate whether genes from the gene-based analysis are represented in conserved regions. Conserved regions are genes which are evolutionary constrained, it is expected that no enrichment will be observed in the AD GWAS data since AD is a post-reproductive onset disorder [77].

- Extend this analysis to consider pathways, which have previously been found to be associated with AD [23][24], and assess whether a similar association is shown using MAGMA software compared to ALIGATOR [61].

3.2 Materials and Methods

The MAGMA gene-based and pathway analysis was run in both the GERAD [19] and IGAP stage 1 [20] data, discussed in Section 2.1. The GERAD data contains raw genotype data, and as such, MAGMA using the PCA approach was used on this data (MAGMA-PCA). The IGAP data consists of summary statistics only, MAGMA has the option to carry out set-based analysis on summary statistics using Brown's method [52] (MAGMA-SUMMARY), and therefore, this approach was used in this data. MAGMA-SUMMARY utilises a mean χ^2 approach for the summary statistics. The mean χ^2 statistic is calculated for the SNPs within a gene and is compared to a known approximation of the sampling distribution. The use of a known sampling distribution is what enables the method to be computationally efficient.

Two separate analyses are considered on the IGAP data, the first using summary statistics for the SNPs in GERAD only and the second using all stage 1 SNPs. Prior to the gene-based analysis, SNP summary statistics for the whole IGAP data were adjusted for the genomic control parameter, λ=1.087, as reported in [21].

3.2.1 Gene-Based Analysis

SNPs were assigned to genes using GENCODE (v19) gene models [78]. Only genes with known gene status and those marked as protein coding were used. SNPs that belong to multiple genes were assigned to all possible genes, where SNPs were within genes only. SNPs were assigned to 14,607 unique genes.

Since only summary statistics were available in the IGAP data, the LD between SNPs cannot be estimated directly from the data. Therefore, the LD was estimated using the 1000 Genomes Project (1000G) data [79]. The 1000G data is a haplotype reference set which can be used by genetic researchers.

MAGMA provides a gene-based p-value for the annotated genes. The number of genes which are gene-wide significant ($p < 2.5 \times 10^{-6}$) were determined. Of these significant

genes, the number of genes which were not identified by SNP regions in the IGAP data were determined. The SNP region is defined as ± 500kilobases (kb) of a single genome-wide significant SNP ($p < 5 \times 10^{-8}$).

The gene-based results using GERAD and IGAP considering only SNPs in GERAD data are directly comparable since the analyses have the same number of SNPs, but IGAP includes a larger number of subjects. The gene-based analysis from the IGAP stage 1 data includes a larger number of SNPs so is not directly comparable with the other results. A gene-based analysis in the combined IGAP stage 1 and 2 data has previously been published and therefore is not repeated here [21].

The MAGMA analysis was repeated where SNPs within a window of a gene were assigned to the gene. A window was used which is 35kb upstream and 10kb downstream of the gene, since transcriptional regulatory elements are likely to be contained within these intervals and there may be merit in capturing the variation within these regions [76]. The annotation and MAGMA analysis was repeated using genes with this window.

The final aim was to determine whether genes identified were in conserved regions. Both for genes that are evolutionary constrained, i.e, that are less likely to harbour variants of strong effect, probably due to functional importance, and for genes in Conserved Noncoding Sequences (CNS) which are less prone to variation and thus thought to play a role in the regulation of their neighbouring genes [80]. Genes which are subject to strong selection against various mutations were determined using the Exome Aggregation Consortium (ExAC) which contains high quality exome sequence data for 60,706 subjects [81]. This database contains a list of genes, so the number of genes from our analysis which reside in this list were determined. The CNS were defined from [82], this data contains genomic locations for these regions. Of the genes from the gene-based analysis, the number which were in conserved regions and the number of significant genes using a gene-wide significant p-value threshold of 2.5×10^{-6} [83] and a nominal threshold of 0.05 were determined. A chi-squared test was then used to determine whether an association between gene significance and whether the genes are in conserved regions exists. This was then extended to investigate SNPs; again a chi-squared test was used to assess whether an association exists

between the number of SNPs in a conserved region and whether SNPs were significant, using a genome-wide significant p-value threshold of 5×10^{-8} and a nominal threshold of 0.05. A nominal threshold was used in the case where small numbers were observed in conserved regions. When cell counts were small, a more robust Fisher's exact test was used instead of a chi-squared test. A chi-squared test assumes that genes are independent, of course, this is not, in general, the case, and therefore results should be interpreted with caution.

3.2.2 Pathway Analysis

The eight pathways discovered as being associated with AD [23][24] were used to determine whether the same pathways are found to be associated with disease using the MAGMA pathway approach.

The pathway analysis in MAGMA builds upon the gene-based analysis. It uses the p-values for the association of a gene to AD and converts them to Z-values. The self-contained test uses the Z-values for the genes in the pathway as the dependent variable in an intercept only regression model. If the intercept is non-zero, this suggests the pathway is associated with disease. The competitive test uses the Z-values for all genes as the dependent variable in a regression model, whilst using a binary indicator, which is equal to one if the gene is in the pathway and zero otherwise, as an independent variable. The pathway tests in MAGMA correct for the effects of gene size and gene density [84].

3.3 Results

3.3.1 Gene-Based Analysis

3.3.1.1 GERAD Results

The results of the gene-based analyses are summarised in Figures 3.1-3.6.

Figure 3.1 shows the results from the MAGMA-PCA gene-based analysis in the GERAD data. The top box shows the total number of genes in the analysis (N_{genes}=14,606), the middle-left box shows the number of genes which are not gene-wide significant (N_{genes}=14,603) and the middle-right shows the number of genes which reach gene-wide significance (N_{genes}=3). The bottom boxes show which of these gene-wide significant genes are identified by genome-wide significant SNPs. The bottom-left box shows the number of gene-wide significant genes which are identified by genome-wide significant SNPs (N_{genes}=3) and the bottom-right box shows the number of gene-wide significant genes which are not identified by genome-wide significant SNPs (N_{genes}=0). All flow diagrams are organised in the same way.

There are three genes which are gene-wide significant, however, these are all close to a genome-wide significant SNP and are explained by the large effect of *APOE*. The results for the MAGMA-PCA gene-based analysis in GERAD data with a window 35kb upstream and 10kb downstream of the gene are shown in Figure 3.2. The use of the window enables a larger number of gene-wide significant genes to be determined, therefore, suggesting a benefit of using a flanking region. However, again, all these genes are explained by genome-wide significant SNPs.

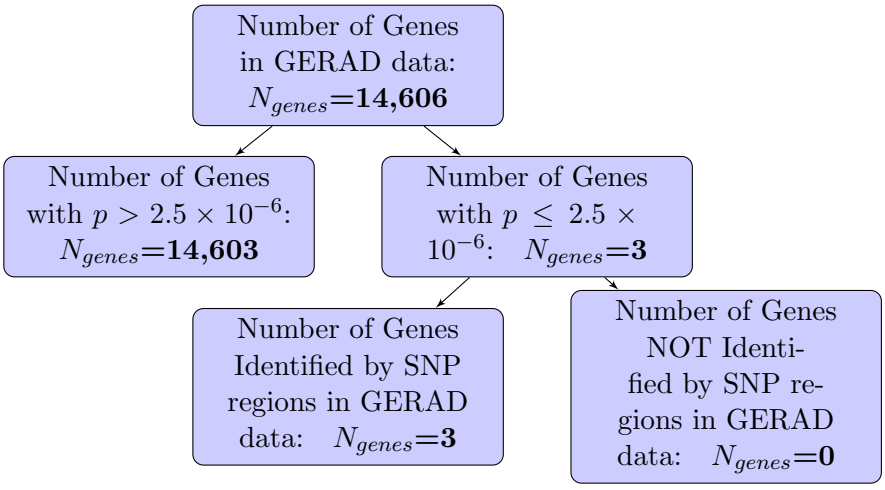


Figure 3.1: MAGMA-PCA Analysis in GERAD Data

Since, no genes are found using the single SNP analysis in GERAD data, it is also true that no additional genes are determined when comparing the GERAD gene-based analysis to the single SNP analysis in the IGAP data (considering only GERAD SNPs). Therefore,

from this analysis, it cannot be concluded that gene-based analyses in smaller data have greater power than single SNP analyses in larger data based on AD genotype data with N=13164.

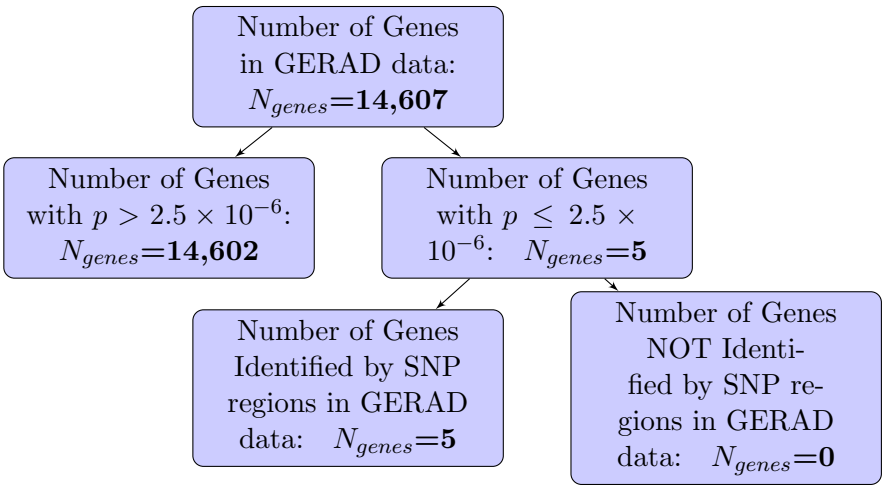


Figure 3.2: MAGMA-PCA Analysis in GERAD Data with a Gene Window

3.3.1.2 IGAP (GERAD SNPs only) Results

The results for the gene-based analysis in IGAP data considering only GERAD SNPs are seen in Figures 3.3 and 3.4, the flow diagrams are organised as in Section 3.3.1.1. The MAGMA-SUMMARY [55] gene-based analysis in IGAP data considering GERAD SNPs only, using the 1000G data to estimate LD between SNPs, seen in Figure 3.3, produced 14,541 genes, 17 of which were significant at the gene-wide level ($p < 2.5 \times 10^{-6}$). All of these 17 genes were identified by significant SNP regions in the IGAP data, where SNP regions are defined as ±500kb of a single significant SNP ($p < 5 \times 10^{-8}$).

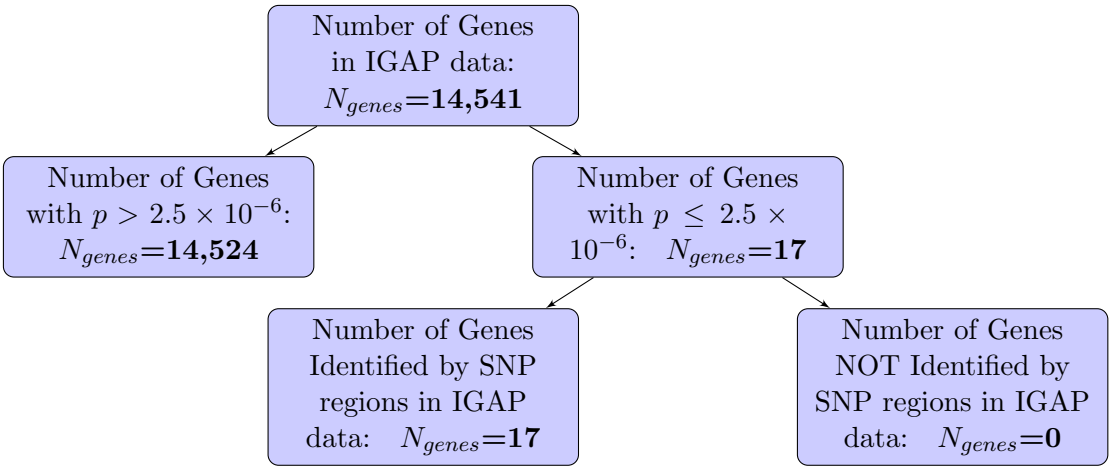


Figure 3.3: MAGMA-SUMMARY Analysis in IGAP Data, GERAD SNPs only

Again, the use of flanking regions results in a larger number of genes reaching gene-wide significance, see Figure 3.4. Although, all of these 24 gene-wide significant SNPs were identified by single SNP analyses.

The conclusions based on this analysis in IGAP using GERAD SNPs only are the same as those based on GERAD data.

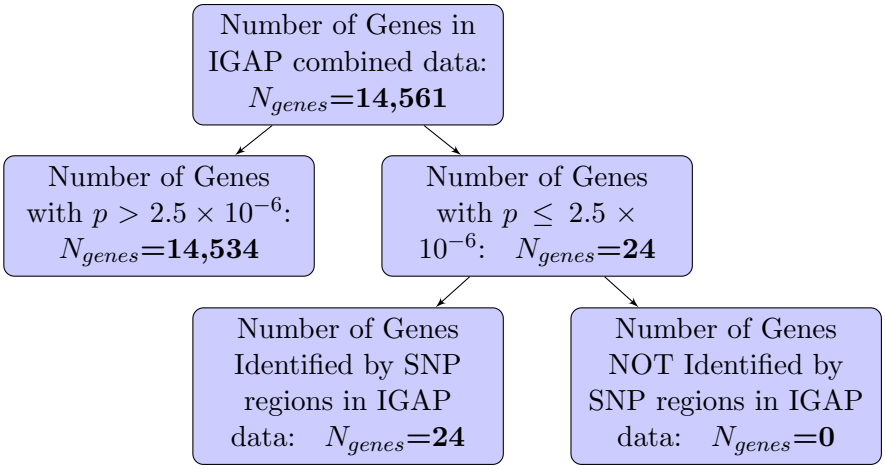


Figure 3.4: MAGMA-SUMMARY Analysis in IGAP Data with a Gene Window, GERAD SNPs only

3.3.1.3 IGAP Stage 1 (All SNPs) Results

The MAGMA-SUMMARY analysis was rerun using all SNPs in the stage 1 IGAP data. The 1000G was used to estimate LD between SNPs. The results can be seen in Figure

3.5. For this analysis, 28 gene-wide significant genes ($p < 2.5 \times 10^{-6}$) were determined, 3 of which were not identified by genome-wide significant SNP ($p < 5 \times 10^{-8}$) regions in IGAP data. These genes were *HBEGF*, *SLC39A13*, and *FLJ00418*.

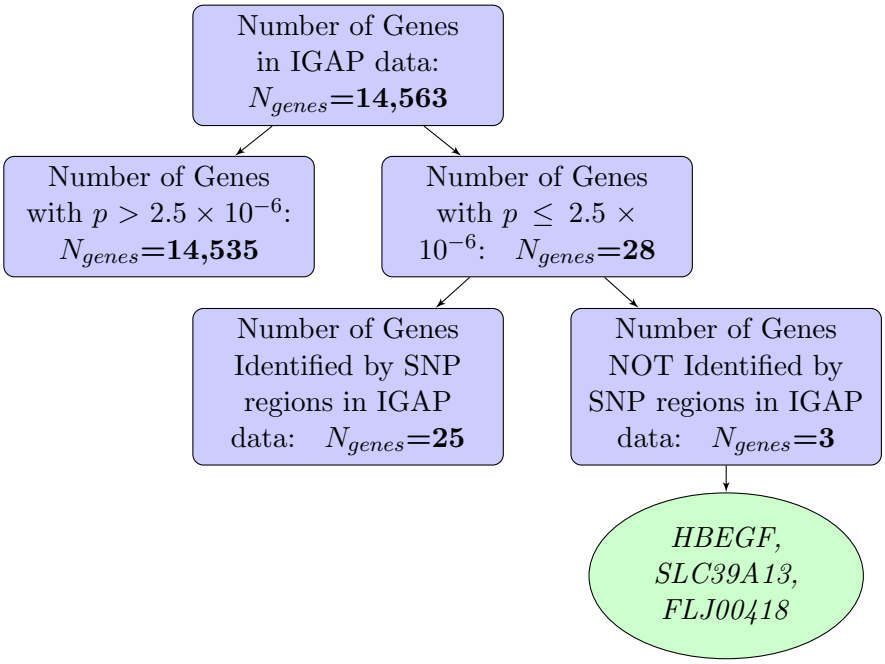


Figure 3.5: MAGMA-SUMMARY Analysis in IGAP Stage 1 Data

The MAGMA-SUMMARY analysis with a gene window of 35kb upstream and 10kb downstream finds a larger number of gene-wide significant SNPs, 2 of which are not identified by single SNP analyses; these genes are *HBEGF* and *SLC39A13*. The *SLC39A13* gene was previously determined by a gene-based analysis in the combined IGAP stage 1 and 2 data [21]. Additionally, the *HBEGF* gene was identified in [21] study using IGAP stage 1 data but this was not replicated in IGAP stage 2 data.

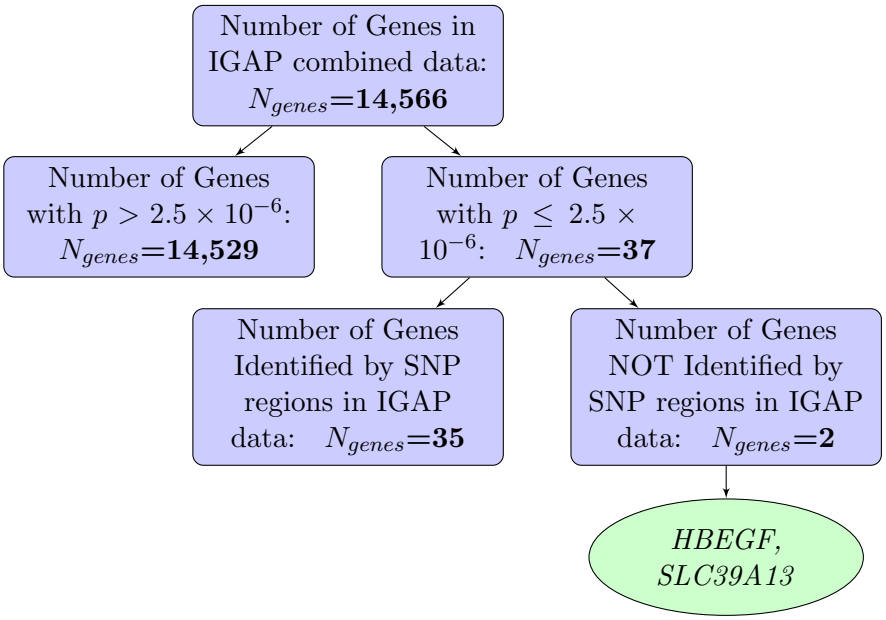


Figure 3.6: MAGMA-SUMMARY Analysis in IGAP Stage 1 Data with a Gene Window

Two genes were determined in this analysis in the IGAP stage 1 data with all SNPs using a gene window which have not previously been determined; *HBEGF* and *FLJ00418*. Although, in the analysis by [21], the effect of these genes were cancelled out when the stage 1 and 2 IGAP results were meta-analysed together.

From these gene-based analyses, it is seen that by combining the effect of SNPs in the IGAP stage 1 data including all SNPs it is possible to attain more power than single SNP analyses. In the smaller sets, GERAD and IGAP only including GERAD SNPs, no improvement in power over single SNP analyses is observed by using gene-based analyses.

Since there was no power gained by gene-based analysis in GERAD data compared to single SNP analysis for AD data, there is no evidence that the gene-based analysis would increase power even in a smaller dataset.

All analyses show that using flanking regions for the gene-based analysis improves power, since the number of gene-wide significant genes increases when the flanking region is used, indicating the potential variation captured by using these regions.

3.3.1.4 Conserved Regions

Loss of Function (LoF) genes from The Exome Aggregation Consortium (ExAC)

The GWAS data were interrogated for genes that are evolutionary constrained, i.e, that are less likely to harbour variants of strong effect, probably due to functional importance. The Exome Aggregation Consortium (ExAC) defines constrained genes using a pLI metric [81]. In effect, Loss of Function (LoF) intolerant genes (pLI ≥ 0.9) harbour considerably less protein-truncated variants than expected. Interestingly, the ExAC study shows that the most highly constrained genes are enriched in GWAS loci. As LoF genes are expected to carry essential functions, however, the common variants within these genes are not expected to have a strong effect, as observed with the relatively low OR associated with GWAS loci.

Of the 14,563 genes identified from the IGAP stage 1 data, Table 3.1 show the number of genes in and out of LoF regions and which of these are above or below the p-value threshold.

Table 3.1: Number of LoF Genes from the Exome Aggregation Consortium

	$p \leq 0.05$	$p > 0.05$
in LoF	251	2496
out LoF	1004	10812

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	7	2740
out LoF	29	11787

(ii) $p \leq 2.5 \times 10^{-6}$

No enrichment is observed for LoF intolerant genes in our AD GWAS dataset for either the nominal p-value threshold or gene-wide p-value threshold (p=0.2986 and p=0.834 (Fisher's exact) respectively). This is perhaps not surprising as AD is a post-reproductive onset disorder and would therefore not be strongly selected against [77]. One possible explanation is "antagonistic pleiotropy" in which genes that exert a beneficial effect (early in life, up to the age of reproduction) become detrimental later in life [77][85]. The immune response system is one such example [85].

Conserved Noncoding Sequences (CNS)

CNS are regions of the genome that are evolutionary constrained, less prone to variation and thus thought to play a role in the regulation of their neighbouring genes [80].

Table 3.2 shows the number of genes in the CNS at each p-value threshold, 0.05 and 2.5×10^{-6} respectively. Since the cell counts are low for both tables, a Fisher's exact test is used rather than a chi-squared test.

Table 3.2: Number of Genes in Conserved Noncoding Sequences

	$p \leq 0.05$	$p > 0.05$
in LoF	3	24
out LoF	1252	13284

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	0	27
out LoF	36	14500

(ii) $p \leq 2.5 \times 10^{-6}$

Again, no enrichment of CNS in the IGAP dataset is observed for either the nominal or gene-wide p-value threshold (p=0.5025 and p=1 respectively, both determined using Fisher's exact test).

3.3.2 Pathway Analysis

A pathway analysis using the MAGMA software was undertaken in both the GERAD and IGAP data, the eight pathways analysed are those which were previously determined as being associated with AD [23]. MAGMA produces both a self-contained (P_{sc}) and a competitive (P_c) test of association. A self-contained test determines whether the pathway is associated with disease, whereas a competitive test determines whether the association of the pathway is in addition to the baseline level of association [84]. Other gene-set approaches, such as ALIGATOR define a competitive test somewhat differently, by assessing whether a gene-set is more strongly associated compared to a permutation set of pathways containing randomly selected genes. Pathways are classed as being associated if the p-value is less than 1.56×10^{-3}; this was determined using a Bonferroni correction ($0.05 \div (8 \text{ pathways} \times 4)$) [71].

Table 3.3: MAGMA Pathway Results in GERAD data

Pathway	Number of Genes	MAGMA-PCA in GERAD				MAGMA-SUMMARY in GERAD			
		Beta	SE	P_c	P_{sc}	Beta	SE	P_c	P_{sc}
Immune response	728	0.00447	0.0304	0.2497	0.0073	0.00143	0.0295	0.4119	0.6431
Regulation of endocytosis	186	0.00666	0.0557	0.1431	0.1354	0.00383	0.0581	0.2782	0.7480
Cholesterol transport	41	0.00311	0.124	0.3179	0.0473	0.00491	0.125	0.2283	0.0749
Hematopoietic cell lineage	64	-0.0111	0.103	0.5646	0.1177	0.00636	0.1	0.1691	0.0627
Proteasome-ubiquitin activity	269	0.00895	0.0449	0.0688	0.0471	0.00646	0.0452	0.1439	0.5256
Reactome hemostasis	385	-0.0107	0.04	0.9516	0.1685	-0.00811	0.0397	0.8992	0.7828
Clathrin	381	-0.00842	0.0411	0.9006	0.5128	-0.00412	0.04	0.7409	0.8852
Protein folding	146	-0.00851	0.0611	0.9192	0.8540	-0.00738	0.062	0.8844	0.9702

Table 3.3 shows the pathway analysis results in the GERAD data. The pathway results using the gene p-values from both MAGMA-PCA and MAGMA-SUMMARY approaches in the GERAD data are presented. It is seen that for both the self-contained and competitive tests, no pathways have evidence of an association. As expected, the majority of the competitive p-values are higher than the self-contained p-values, although this pattern is less consistent when the MAGMA-SUMMARY approach was used. For some of the pathways, the p-values vary between those using the MAGMA-PCA approach and the MAGMA-SUMMARY approach.

Table 3.4: MAGMA and ALIGATOR Pathway Results in IGAP stage 1 data

Pathway	Number of Genes	Beta	SE	P_c	P_{sc}	ALIGATOR p-value
Immune response	722	0.0071	0.0326	0.1575	0.0133	0.00266
Regulation of endocytosis	186	0.00527	0.0653	0.2353	0.1876	0.0002
Cholesterol transport	41	0.0128	0.13	0.0320	0.0004	0.00024
Hematopoietic cell lineage	64	0.00967	0.108	0.0870	0.0054	0.00007
Proteasome-ubiquitin activity	269	-0.00565	0.0512	0.7938	0.0934	0.00929
Reactome hemostasis	384	0.00187	0.0451	0.3978	0.2569	0.00785
Clathrin	378	0.0189	0.0459	0.0049	0.0368	0.00038
Protein folding	146	0.00859	0.0694	0.1069	0.0202	0.00634

Table 3.4 shows the same analysis in the IGAP stage 1 data, with the additional ALIGATOR [61] p-value [23][24] presented for comparison. IGAP gains power by having a larger number of subjects, despite this, no pathways reach significance based on the competitive test, but the cholesterol transport pathway now reaches significance for a self contained association with AD. The p-values found from the ALIGATOR analysis are consistently lower than those found using the MAGMA approach.

The results differ between MAGMA and ALIGATOR, this is likely due to the difference between the two approaches. Table 3.5 shows the differences in the number of genes included in each pathway between the MAGMA and ALIGATOR approaches; ALIGATOR has a much smaller number of genes in each pathway compared to MAGMA. ALIGATOR only selects genes which contain at least one SNP with a p-value below a set p-value threshold and the number of significant genes in a GO category are compared to 5000 replicate gene lists [61]. MAGMA pathways will contain more noise due to the inclusion of all SNPs in a pathway, including those which may have no evidence of an association with disease. Due to the difference in approaches, the alternative hypothesis being tested differs between ALIGATOR and MAGMA; ALIGATOR is testing whether the gene-set contains a larger than expected number of significant genes, whereas MAGMA tests whether the

combined effect of genes in a pathway is associated with disease.

Table 3.5: Number of Genes in Each Pathway from the MAGMA and ALIGATOR Approaches

Pathway	Number of Genes in MAGMA	Number of Significant Genes in ALIGATOR
Immune response	722	5
Regulation of endocytosis	186	14
Cholesterol transport	41	8
Hematopoietic cell lineage	64	11
Proteasome-ubiquitin activity	269	5
Reactome hemostasis	384	25
Clathrin	378	7
Protein folding	146	12

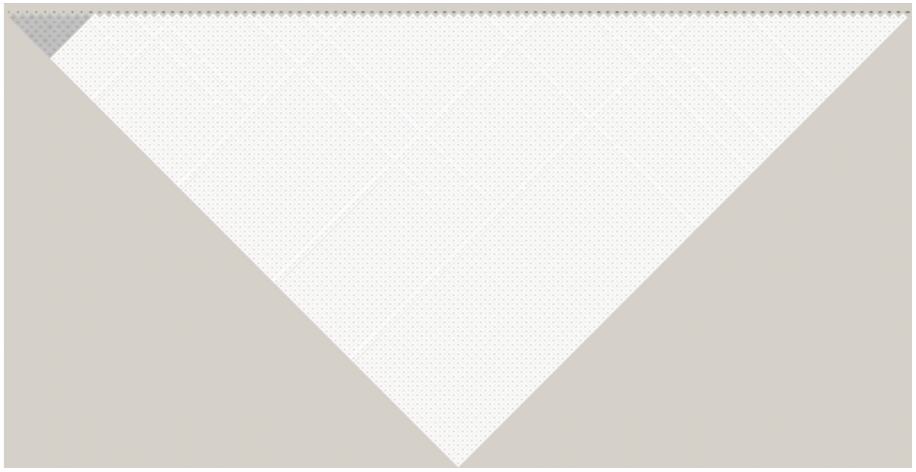
3.3.3 Comparison of MAGMA Settings

A slight difference in pathway results is observed depending on whether MAGMA-PCA or MAGMA-SUMMARY is used. Therefore, the difference between results from these two different approaches is investigated.

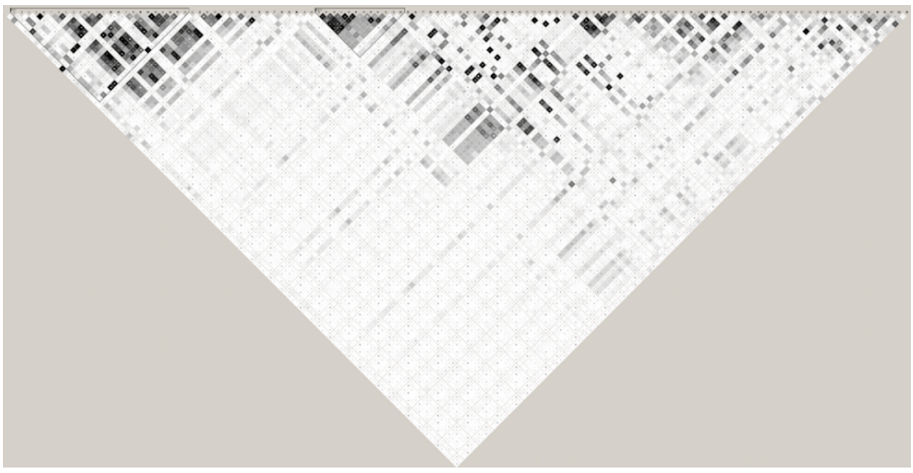
The purpose of this section is to report findings showing issues with the power of set-based analysis using summary statistics, and to provide a theoretical explanation for them. In particular, the power of the set-based analysis using MAGMA on summary statistics can appear to be substantially larger than the power of MAGMA-PCA or Multivariate Regression on the exact same genotype data. The PCA approach was reported to give a clear advantage in power [55] over other methods implemented in MAGMA, including the mean of the χ^2 statistic for the SNPs in a gene [52], and the top χ^2 statistic among the SNPs in a gene. In the latter approach, an adaptive permutation procedure is used to obtain an empirical gene p-value. The authors emphasize *"...that although MAGMA can perform analysis of summary statistics, raw data analysis should always be preferred*

if possible" [55]. The findings reported below indicate that analysis of summary statistics does not provide a reliable reference for comparisons of test power in the presence of correlation.

A data set of 100 SNPs is simulated with MAF=0.2. Of 100 SNPs, ten were in LD with $r^2 = 0.2$ and 0.8, and the remaining SNPs were independent (see LD plot in Figure 3.7a for $r^2 = 0.2$). All SNPs were in HWE. The data were simulated for 10,000 cases and 10,000 controls (20,000 subjects in total). Simulations are repeated 500 times. To test the type I error, effect sizes are fixed at OR=1 for all SNPs. For the power estimation, effect sizes are defined as OR=1.1 for SNPs in the LD block.



(a) Simple LD Structure



(b) Real LD Structure

Figure 3.7: LD Plots for Two Simulated Scenarios

MAGMA is run on these scenarios using different settings, in particular:

- (PCA) - using all available genotypes

  MAGMA syntax: *magma --bfile all.data --gene-annot gene.loc.annot*

- (SUMMARY) - using summary statistics generated by PLINK [50][51] with logistic regression

  MAGMA syntax: *magma --bfile all.data --pval all.data.assoc.logistic N=20000 --gene-annot gene.loc.annot*

- (PART) - using summary statistics of one part of the split data (5,000 cases and 5,000 controls), and deriving LD from the other part of the data.

  MAGMA syntax: *magma --bfile part2 --pval part1.assoc.logistic N=10000 --gene-annot gene.loc.annot*

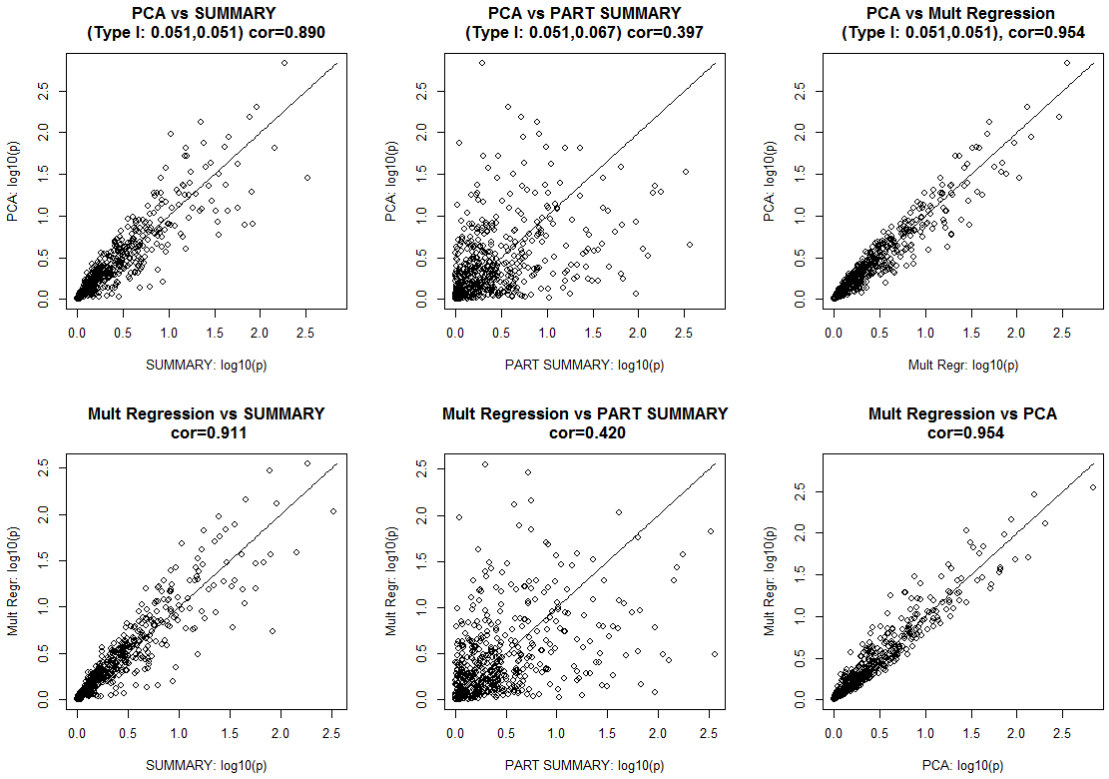
Multivariate regression analysis (Mult Regression) was used on the exact same data using all available genotypes with glm(CasCon.status $\sim$ SNP_1+SNP_2+...+SNP_{99}+SNP_{100}, family=binomial) function in R-statistical software. The SNP-set association p-value was calculated using the Log-Likelihood ratio test p= 1-pchisq(model$null.deviance-model$deviance, df=model$df.null-model$df.residual)

Real LD structure was also investigated, simulating null associations and OR=1.1 in the correlation block (for LD structure see Figure 3.7b).

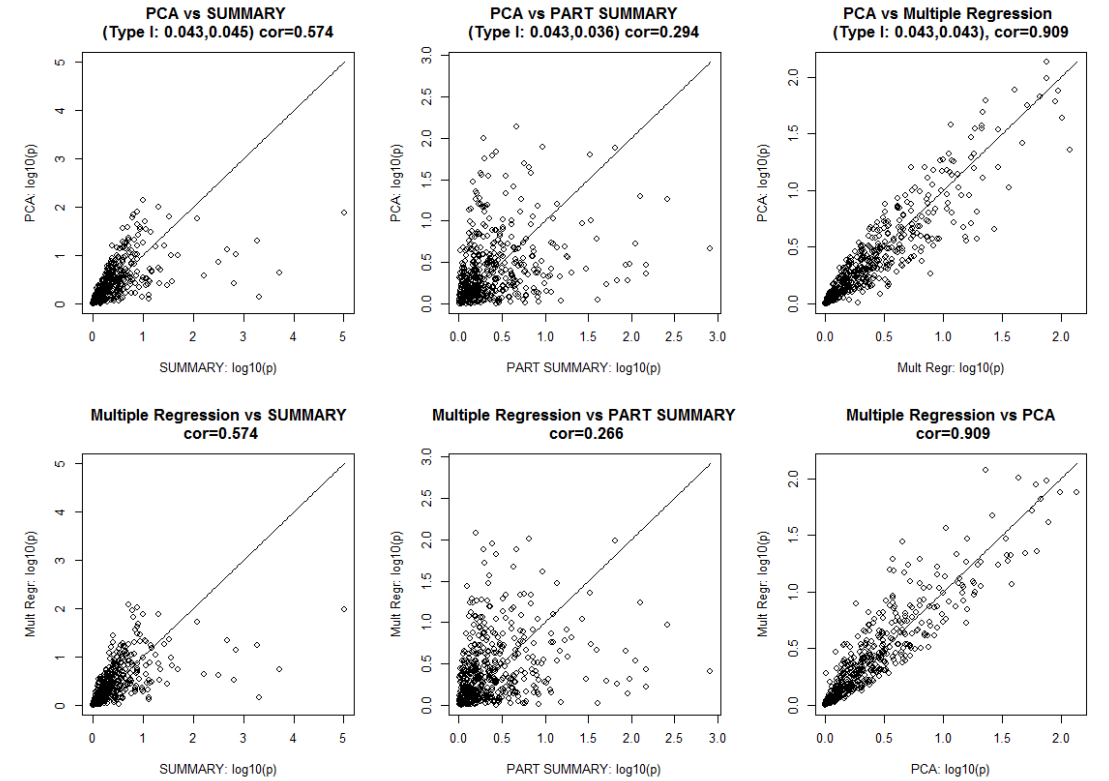
3.3.3.1 Type I error

The type I error was around 5% at 0.05 significance level for all settings: (0.051, 0.051, 0.067, 0.051) when $r^2 = 0.2$ between 10 SNPs in a block, and (0.043, 0.045, 0.036, 0.043) when $r^2 = 0.8$ between 10 SNPs in a block, for (PCA), (SUMMARY), (PART) and (Mult Regression) settings, respectively. The correspondence between p-values under the null hypothesis for all SNPs was quite good (see Figures 3.8a, 3.8b). Reassuringly, the highest correlations between $-log_{10}$(p-values) (0.96 and 0.91, for r^2 in the LD block 0.2 and 0.8, respectively) were observed for PCA and Multiple regression settings. The weakest correlation (0.4, 0.3) was between $-log_{10}$(p-values) generated by (PCA/Mult regression) and (PART) settings. This makes sense as in the (PART) setting only half of the data was used to estimate SNP p-values and the other half was used to determine LD between SNPs. The correlation between (PCA) and summary statistic based settings (SUMMARY) was

0.902 and 0.574, for r^2 in the LD block 0.2 and 0.8, respectively.



(a) Simple LD Structure



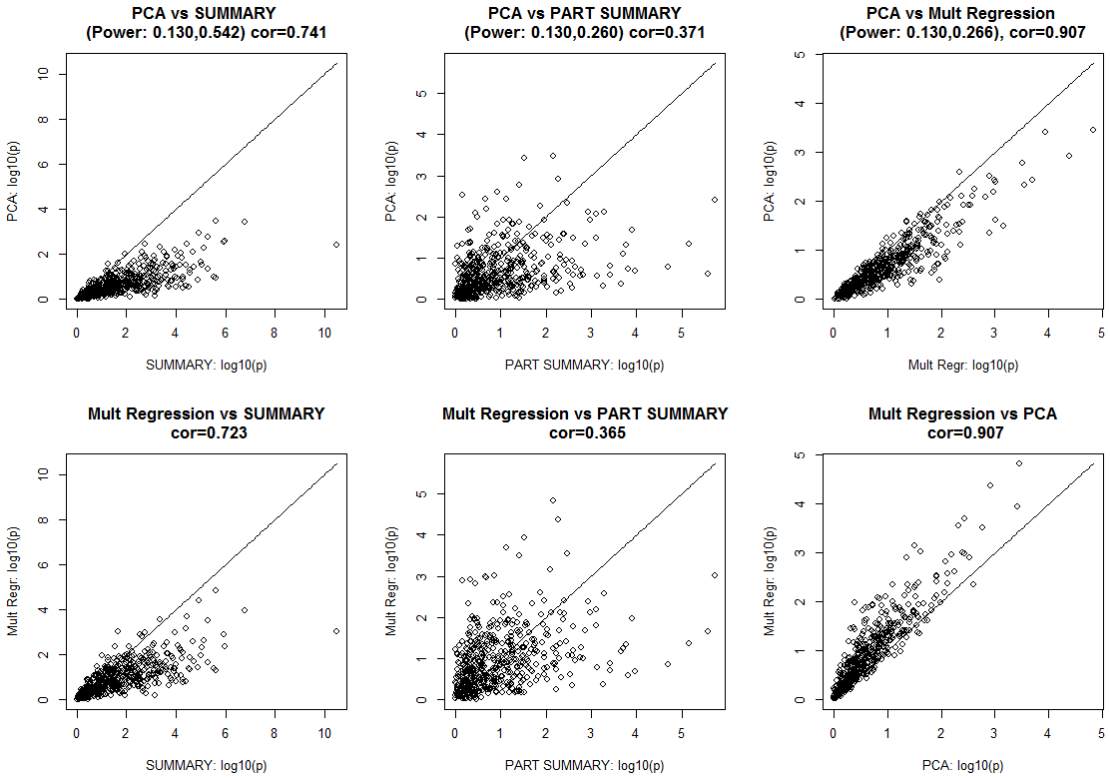
(b) Real LD Structure

Figure 3.8: Scatter plots of $-log_{10}$(p-values) generated with (PCA), (SUMMARY), (PART) and (Mult Regression) settings for 500 simulated sets of 100 SNPs, of which 10 SNPs were in LD. All SNPs association OR=1 (Null Hypothesis).

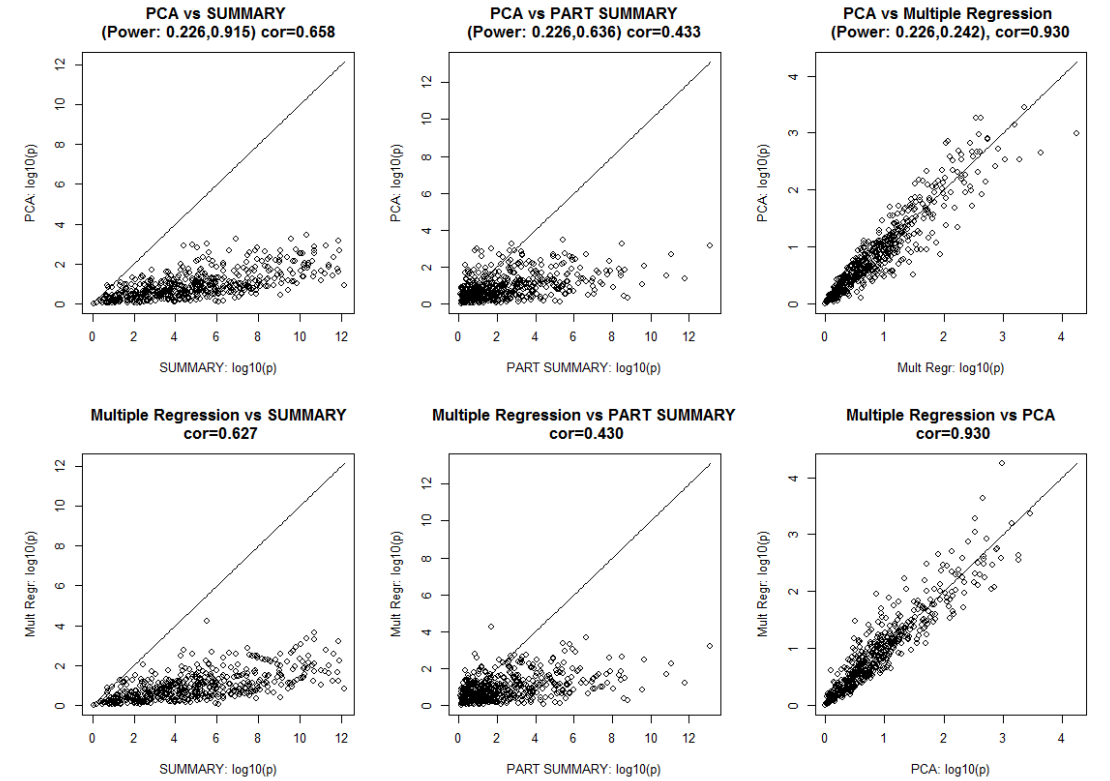
3.3.3.2 Power

Figures 3.9a, 3.9b show scatterplots comparing $-log_{10}$(p-values) of the different analyses, with LD block $r^2 = 0.2$ and $r^2 = 0.8$, respectively. The power for the PCA setting of MAGMA and Mult Regression in R, give similar power estimates with a slight advantage of Mult Regression (compare power 0.119 vs 0.251 and 0.226 vs 0.242 in the right plots in Figures 3.8a, 3.8b, respectively). This is expected as the PCA-setting of MAGMA projects the SNP matrix for a gene onto its PCs, pruning away PCs with very small eigenvalues, and then uses those PCs as predictors for the phenotype in the regression model.

However, Figures 3.9a, 3.9b show consistently higher power for summary stats based analyses over PCA and Multiple regression settings. The weakest correlation (0.4, 0.3) was between $-log_{10}$(p-values) generated by (PCA/Mult regression) and (PART) settings. Note that in the (PART) setting only half of the data were used to estimate SNP p-values and the other half were used to determine LD between SNPs. The correlation between (PCA) and summary statistics based settings (SUMMARY) was 0.902 and 0.574, for r^2 in the LD block 0.2 and 0.8, respectively. These findings show that even when the type I error rate remains correct, the p-value for an individual sample found from summary statistics can differ widely from that found using PCA or Mult Regression in the presence of LD.



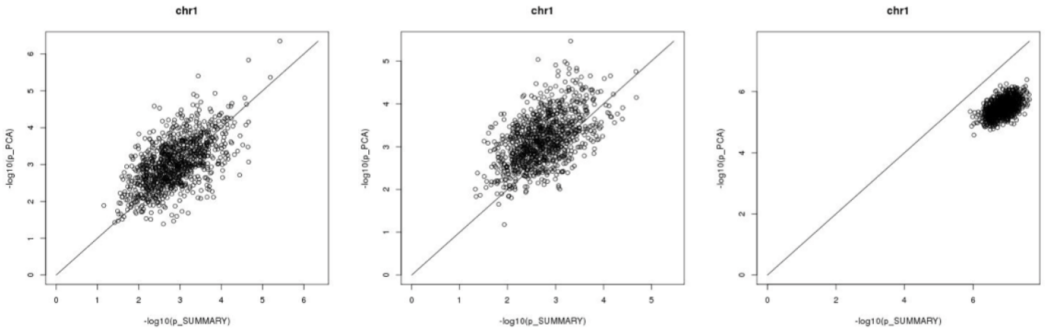
(a) Simple LD Structure



(b) Real LD Structure

Figure 3.9: Scatter plots of $-log_{10}$(p-values) generated with (PCA), (SUMMARY), (PART) and (Mult Regression) settings for 500 simulated sets of 100 SNPs, of which 10 SNPs were in LD. SNPs association ORs=1.1 for SNPs in the LD block, OR=1 otherwise.

These simulated scenarios have a very simple LD structure and effect sizes as compared to real data, and may overemphasise the type I and II errors. Therefore, a real LD structure was also considered, simulating OR=1.1 in the correlation block (for LD structure see Figure 3.7b). Similar patterns of discrepancies were observed between the p-values generated by (PCA) and (SUMMARY) settings of MAGMA (see Figure 3.10).



(a) Null (b) Associated SNP not in LD Block (c) Associated SNP in LD Block 2

Figure 3.10: Comparison of MAGMA Settings in Real Data (Simulation of 115 SNPs from Real Data, with a Proportion of Phenotypes Permuted to Maintain Effect Sizes, Test and Discovery Set N=13,164.) Varying the Position of the Associated SNP.

Test power calculated for Brown's method (MAGMA-SUMMARY) may overestimate or underestimate the actual joint power of the tests, as defined in terms of the overlap of the sampling distributions under the alternate and null hypotheses, depending on the relationship between the mean vector of the alternate hypothesis and the correlation matrix. A situation where the effects are roughly aligned with the correlation, as is commonly the case in genetic data, will lead to an overestimation of the actual test power, which would be found correctly from multivariate regression or Hotelling's T^2 test. This is shown in Figure 3.11 where the isotropic statistic given by Brown's method (circle) is compared to the correct ellipsoidal distribution given by correlated test statistics. Region A in the figure shows points which would be misclassified as not associated with disease and region B shows those which would be classed as associated with disease when in fact they are not.

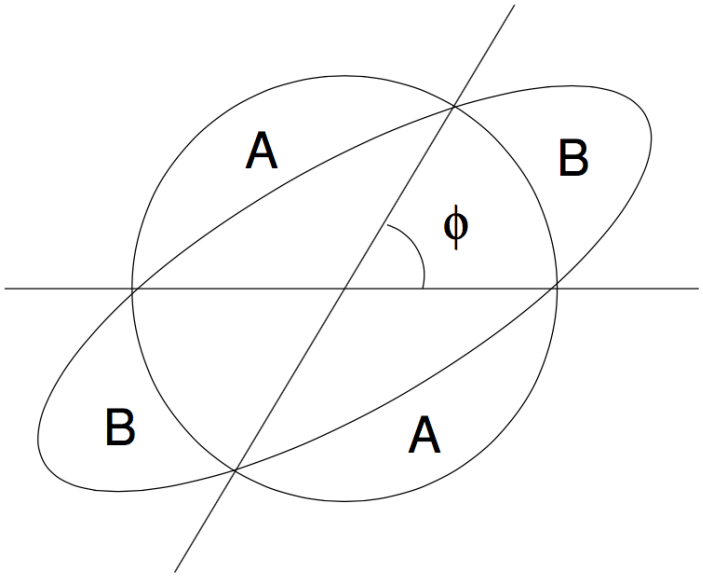


Figure 3.11: Comparison of classification by isotropic combined statistic (circle) with classification according to ellipsoidal distribution of correlated test statistics (ellipse). corr = cos ϕ. Data points in regions A will be misclassified as negative, data points in regions B as positive when the isotropic statistic is used.

In practice, there might be a situation when only summary data are available. In this case it is sensible to prune SNPs for LD before using MAGMA-SUMMARY on summary statistics. MAGMA authors recommend using MAGMA-PCA on individual genotype data where available rather than summary statistic data. When individual genotype data is available, there is no need to LD prune the data when using MAGMA-PCA.

There is no consensus in the field as to whether the inflation seen using a summary statistic approach (such as MAGMA-SUMMARY or Brown's method) in the presense of strong LD gives an advantage in that it is possible to identify significant sets or whether this is a power inflation, as type I error is well controlled using this approach. Brown's method is useful when only summary statistics are available and it may be possible to find a larger number of significant sets in the presence of strong LD. It is a personal opinion that raw genotypes should always be preferable, where available. Although, it is matter of preference and therefore for each researcher to decide which method they prefer.

3.4 Discussion

The first aim of this chapter was to investigate whether gene-based analysis has more power than single SNP analyses by combining the effect of all SNPs in the gene. It was shown in the IGAP stage 1 data including all SNPs, that additional genes may be determined using a gene-based analysis compared to a single SNP analysis. Two genes were identified which have not been found to be previously associated with AD, *HBEGF* and *FLJ00418*, however, these genes were not found in a previous gene-based analysis which meta-analysed stage 1 and 2 of IGAP [21] suggesting a spurious association of some SNPs, which correctly did not pass the genome-wide significance threshold in IGAP stage 2 data.

It was also determined whether it is possible to find equivalent signals using a gene-based analysis in smaller data compared to a single SNP analysis in larger data. In the data used here, the gene-based analysis in GERAD was not powerful enough to find more associated genes than a single SNP analysis in the larger IGAP set.

It was considered whether the analysis could be enhanced by using flanking regions around the gene, since transcriptional regulatory elements are likely to be contained within these intervals and there may be some merit in capturing the variation within these regions [76]. In all analyses considered, the use of a flanking region; 35kb upstream and 10kb downstream, always resulted in a larger number of gene-wide significant genes being found.

It was demonstrated that genes from the gene-based analysis are not represented in conserved regions. There was no enrichment observed for LoF intolerant genes or CNS. As suggested, this is not surprising as AD is a post-reproductive onset disorder and would therefore not be strongly selected against [77]. One possible explanation is "antagonistic pleiotropy" in which genes that exert a beneficial effect become detrimental later in life [77][85]. The immune response system is an example of this [85] and was a pathway previously found to be associated with AD [23][24].

This analysis was extended to consider eight pathways which have previously been found to be associated with AD [23][24], and it was assessed whether a similar association is shown

using MAGMA software compared to ALIGATOR [61]. There is little association for the eight pathways shown in the GERAD data only, and only the cholesterol transport pathway is significant in the IGAP stage 1 data based on a self-contained test. A larger number of pathways reach significance in the self-contained test of association, but this does not withstand the correction for the baseline level of association. The MAGMA pathway analysis results showed a weaker association for all pathways compared to ALIGATOR results. For the pathway analysis in the GERAD data, it was seen that the pathway results vary slightly depending on whether the MAGMA-PCA or MAGMA-SUMMARY approach was used for the gene-based analysis. This is unexpected, since both approaches were applied to the same data.

The differences between the MAGMA-PCA and MAGMA-SUMMARY approaches were compared to each other and to multiple regression using simulated data, both with a simple constructed and real LD structure. In all examples, the type I error is similar for all methods, however, when the associated SNPs reside in the LD block, the power is often higher when using the MAGMA-SUMMARY method compared to the MAGMA-PCA method. This seems strange, since one would expect the raw data to always be the optimal data to use, in fact, the MAGMA authors recommend using the MAGMA-PCA approach where possible. Of course, it may be the case that only summary statistic data are available (i.e. IGAP summary statistics), in which case, it may be sensible to LD prune the data prior to using MAGMA-SUMMARY to avoid potentially spurious results.

Although it has been shown that it is possible to increase power over single SNP analyses by using a gene-based approach, the power increase is minimal and has not determined any novel genes associated with AD. It would be interesting to consider other approaches to gene-based analyses, which may further increase power.

4 Polygenic Risk Score Set-Based Approach

4.1 Introduction

Polygenic Risk Scores (PRSs) are now widely used for a variety of purposes in assessing the genetic liability to disorders or more general phenotypes. These include sample stratification, risk prediction, and the detection of relationships between different subphenotypes (see e.g. [86], [38], and [87], respectively). The PRS method can also be adapted to partition the polygenic risk based on meaningful SNP sets, such as genes or biological pathways, and to determine whether a set of SNPs, weighted with their individual genetic risk effects, is associated at the whole-genome or set-specific level. In contrast to set analysis which aims to analyse the joint association of SNPs with a single phenotype, PRS aims to assess the genetic liability to some phenotype on the basis of the polygenic risk for the same or a different phenotype estimated from independent data.

As discussed in Chapter 3, set-based analysis is an attractive alternative to single SNP analyses since the combined effect of SNPs within a set may be captured. Single SNP analyses are often underpowered due to the small effect sizes of individual SNPs. Set-based analysis considers the combined effect of all SNPs within the set, which would be expected to have a larger effect size and thus be easier to detect. The results in Chapter 3 did not show very large power increases between set-based analysis and single SNP analyses, therefore, this chapter aims to further increase the power of the set-based analysis.

PRS analysis can be considered as set-based analysis when a set includes all SNPs in the whole genome. PRSs provide a method for combining information from individual SNPs

into a single measure of risk allele burden. In their most widely used form, PRSs have been applied to genome-wide SNP data where they can capture a useful fraction of genetic liability to polygenic traits. PRSs can also be used as genome-wide predictors of affected status [35][37][38]. It seems reasonable that the basic principles of polygenic score analysis can also be applied to individual genes, or to gene-set analyses. The motivation for doing so is somewhat different than PRS analyses of genome-wide data; which is to detect associations with potentially biologically informative features rather than predict case-control status or trait liability captured. As for genome-wide analyses, genes or gene-set PRS could be used to predict affected status, or to estimate the gene-specific or pathway-specific SNP liability captured by GWAS. However, for polygenic disorders where risk is dispersed across hundreds of genes and multiple gene-sets, self-evidently, gene-specific or pathway-specific SNP liability to a phenotype will be lower than the liability captured by genome-wide data, and accordingly, such tests will afford less case-control discriminatory power. However, if the disease is homogeneous and is caused by different biological processes, then the biologically relevant pathway may potentially predict a subphenotype of disease better than the whole genome PRS. Risk scores for each individual per set can be used to stratify individuals for follow-up studies and prioritise genes for further functional studies.

Other set-based methods, discussed in Section 1.3.3, have advantages and limitations, however, these methods do not incorporate the effect sizes from external data whilst maintaining a self-contained test of association in the individual genotype data. Set-based methods using individual SNP p-values are also able to improve power by incorporating external data available from previous studies using meta-analysis, although evidence of an association in this case could result solely from the external data.

The application of PRSs to a set-based framework which informs the analysis with previously reported effect sizes of a SNPs association to disease is suggested. The PRS is calculated per person per set, and the overall set effect is computed using logistic regression.

4.1.1 Objectives

The aims of this chapter are to:

- Propose the use of PRS as a set-based approach.
- Evaluate the $\text{PRS}_{set-based}$ method in simulated data, both with constructed and real LD structure.
- Compare the type I error and power for the $\text{PRS}_{set-based}$ method to other set-based approaches, namely; MAGMA[55], Fisher's method [45] for combining p-values and Simes' method [46] which finds the smallest adjusted p-value.

4.2 Materials and Methods

4.2.1 Polygenic Risk Scores

In order to apply the use of PRSs [37] for set-based analysis, PRSs were calculated for each set of SNPs and each individual using the SNP effect size ($\beta = log_e(OR)$) from a discovery dataset and the number of risk alleles from an independent test dataset.

PRS is a weighted score, where the number of risk alleles an individual has for a SNP is multiplied by the SNP effect size from an external discovery set, and this is summed across all SNPs in the set, see Section 2.2.4 for more details. This approach assumes independence between the SNPs included in the score.

For all analyses below, missing genotypes in real data were processed by estimating the missing genotypes using MAF. The SNP genotype was estimated as 2× MAF, this is the same approach used in the PLINK software [50].

Once the PRSs have been generated, a set-based p-value is calculated using logistic regression, adjusting for population covariates. The logistic regression model can be seen in

Equation 4.1.

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 PRS + \text{population covariates} \tag{4.1}$$

where π is the risk of disease, β_0 is the intercept and β_1 is the effect size for the association of PRS with disease, often interpreted as the log(OR). Similarly, for a continuous phenotype, linear regression can be used to determine the set-based p-value.

The strength of association between the set-based PRS and disease is given by $p_{\_PRS}$, which is the p-value for the β_1 coefficient in the logistic regression model.

The application of PRSs as a set-based method was compared to MAGMA [55] gene analysis, Fisher's method [45][53] of combining p-values and Sime's set-based method [46].

4.2.2 MAGMA

MAGMA [55] is a regression based approach (see Chapter 3) which fully accounts for LD between SNPs, using the PCs of the SNP correlation matrix, see Section 2.2.3 for further details. This MAGMA-PCA method is implemented in both the test set only and the discovery and test set combined, since this utilises all data which is used to calculate the PRS, although this method requires the raw genotypes for both datasets.

MAGMA also has the functionality to calculate set-based p-values using summary statistics (MAGMA-SUMMARY). In order to allow for the fairest comparison between this and the PRS method, by utilising all available data, the discovery set summary statistics are used for the gene-based analysis and the test set is used to estimate LD between SNPs and also on the summary statistics for the combined test and discovery sets.

4.2.3 Fisher's Method

Fisher's method [45][53] is a method of combining the p-values of SNPs within a set to determine the evidence against a common null hypothesis [70]. It obtains an overall p-value

for each set from a group of p-values which are independent tests of the same hypothesis, see Section 2.2.1 for the equation to compute Fisher's method.

Since this method is calculated in the test set summary statistics only, no information regarding the direction of the effect of associated SNPs is utilised, however, the $PRS_{set-based}$ method is able to use this information. Therefore, to allow for a fairer comparison, a one-sided p-value is calculated for SNPs associated with disease ($p_{\text{one-sided}} = \frac{p_{\text{two-sided}}}{2}$). These adjusted p-values are then used to calculate Fisher's set-based p-value.

4.2.4 Simes' Method

Simes' [46], like Fisher's, is a set-based method accounting for the number of SNPs within the set, however, it also attempts to correct for LD between SNPs in the set. The null hypothesis of this method is that no SNP within the set is associated with disease, and the alternative hypothesis is that at least one SNP within the set is associated with disease [44]. The Simes' set-based p-value is calculated using Equation 2.2. Simes removes the issue of LD by considering whether at least one SNP in the set is associated with disease, rather than combining the effect of all SNPs. Details of Simes' method are presented in Section 2.2.2.

When compared to the PRS approach, Simes is also informed by the discovery set by using a one-sided p-value for those SNPs which are associated with disease to ensure this method utilises all data available to the PRS method.

4.2.5 Power Comparison Between Methods

Genotype data for both the test and discovery sets were simulated in order to compare the power between a number of set-based methods. A number of different scenarios were simulated;

- **100 SNPs:** 100 independent SNPs, a percentage of which are associated with disease.

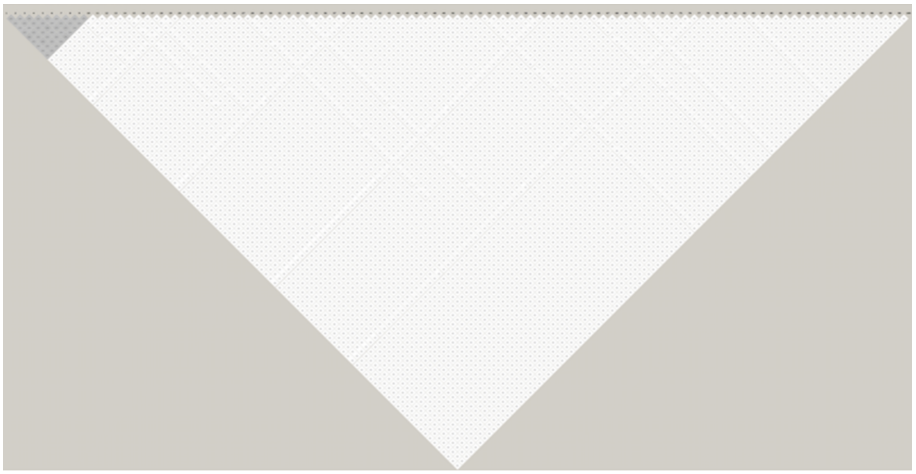
- **Simple LD Block:** 10 SNPs in LD with i) $r^2 = 0.2$ and ii) $r^2 = 0.8$ which are associated with disease, OR=1.1, and 90 independent unassociated SNPs, see Figure 4.1 for LD structure.

- **Complex LD:** 4 LD Blocks of 10 SNPs each, and 60 independent unassociated SNPs. Block 1 has pairwise $r^2 = 0.2$, Block 2 has pairwise $r^2 = 0.4$, Block 3 has pairwise $r^2 = 0.6$, and Block 4 has pairwise $r^2 = 0.8$, all 40 SNPs in LD with OR $\sim$ N(1.02, 0.36) (OR from a Normal Distribution with mean 1.02 and variance 0.36), see Figure 4.2 for LD plot. The mean and variance for the sampled effect sizes are calculated from all SNPs in the IGAP data [20].

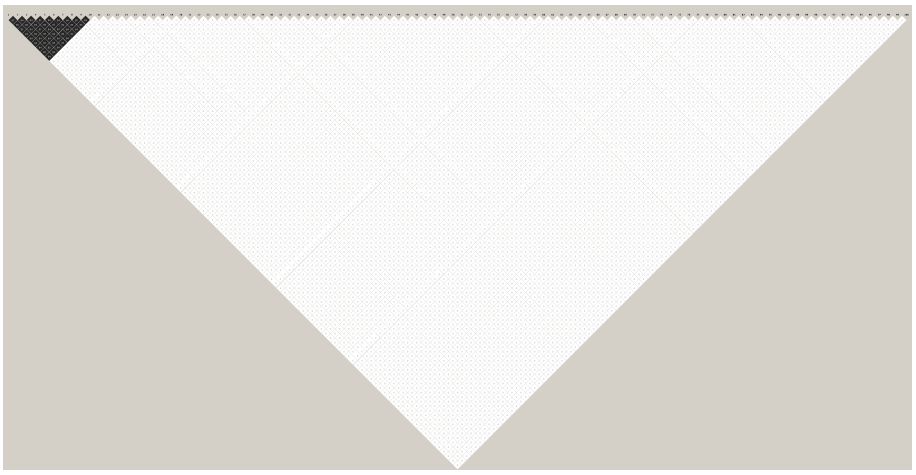
- **Discovery and Test with Different LD Structure:** 10 SNPs in LD with OR$\sim$ N(1.02, 0.36) and 90 independent, unassociated SNPs where test set LD is moderate ($r^2 = 0.6$) and discovery set LD is high ($r^2 = 0.8$), see Figure 4.3 for LD plot.

- **Effect Sizes of Varying Direction:** It is possible that for certain MAFs, SNPs in LD have effects in opposite directions [88]. 10 SNPs with varying LD with ORs with randomly varying direction and 90 independent unassociated SNPs, LD plot shown in Figure 4.4.

- **Real Data Simulations:** 129 SNPs from real AD data, with case control status permuted, while maintaining effect sizes. The real LD structure is shown in Figure 4.5.



(a) $r^2 = 0.2$



(b) $r^2 = 0.8$

Figure 4.1: LD Plot for 100 SNPs in Simple LD Simulations

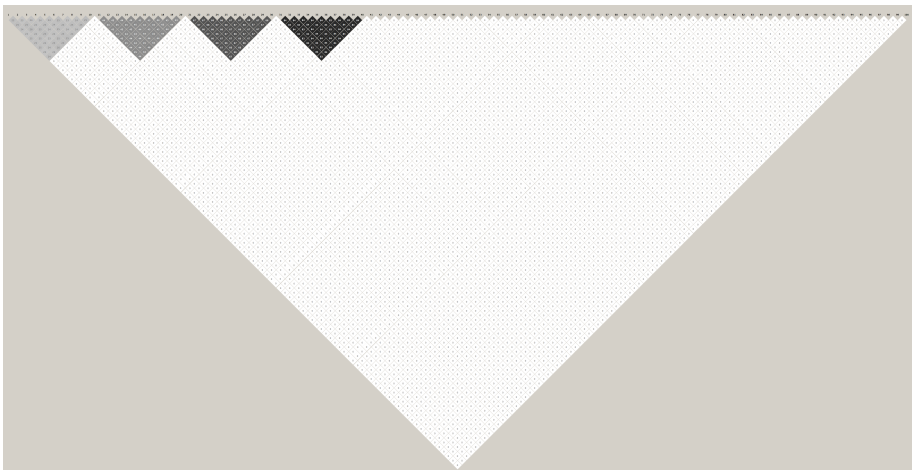
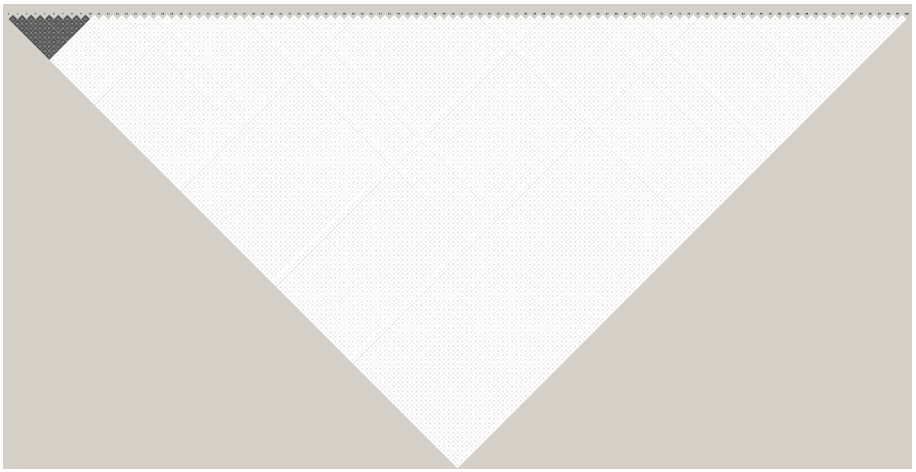
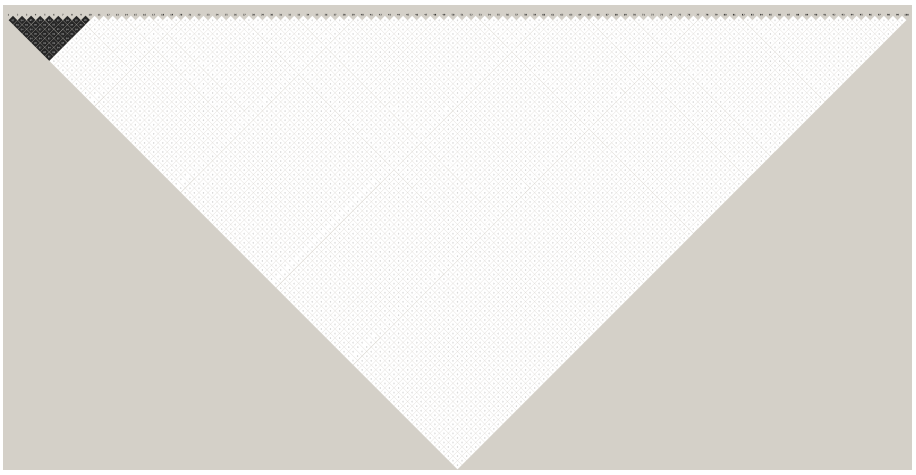


Figure 4.2: LD Plot for 100 SNPs in Complex LD Simulations



(a) Test Set $r^2 = 0.6$



(b) Discovery Set $r^2 = 0.8$

Figure 4.3: LD Plot for 100 SNPs in Discovery and Test with Different LD Structure Simulations

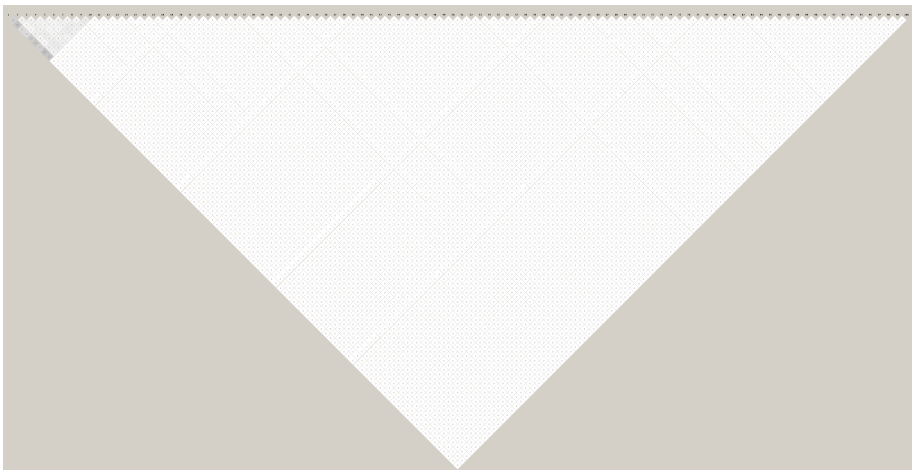


Figure 4.4: LD Plot for 100 SNPs with Varying Effect Sizes Simulation

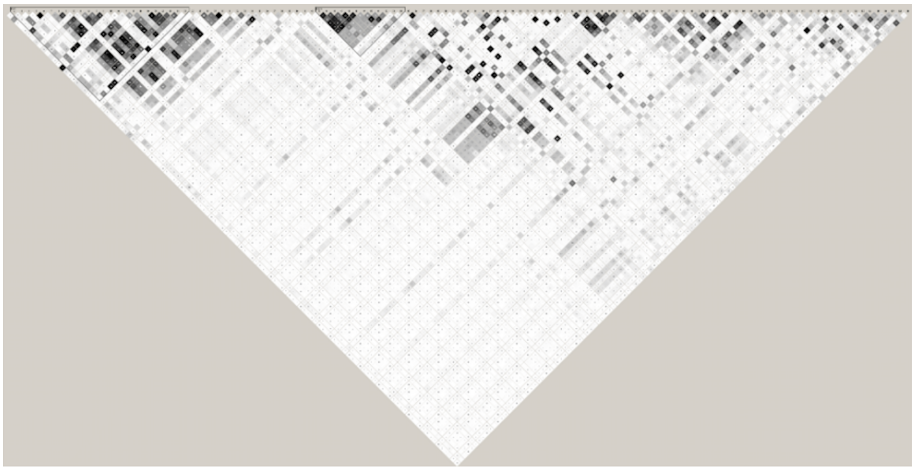


Figure 4.5: LD Plot for Real LD Structure

A total of 1,000 simulations were used for each scenario. The power to detect the association between the set and disease is calculated as the proportion of p-values from the 1,000 simulations which were below a p-value threshold; the p-value thresholds used were p=0.05, 0.01 and 0.001. The power of the $\text{PRS}_{set-based}$ method was compared to the power of MAGMA-PCA (calculated in both the test set only and the combined discovery and test set), MAGMA-SUMMARY in the discovery set with LD estimated from the test set, MAGMA-SUMMARY in test and discovery data combined, and Simes' and Fisher's methods calculated in the test set only.

In all cases, the sample size of the discovery dataset was varied, in order to determine the influence of the discovery set sample size on the $\text{PRS}_{set-based}$ method. Simulations were run with N=10,000, 30,000 and 50,000 for the discovery set. The test dataset contained 10,000 or 30,000 subjects. In both the discovery and test datasets, 30% of the sample size were cases. The case/control ratio was defined to have fewer cases than controls since this is usual in real data, i.e. in AD data (17,008 cases, 37,154 controls) [20] and in SZ data (36,989 cases, 113,075 controls) [35].

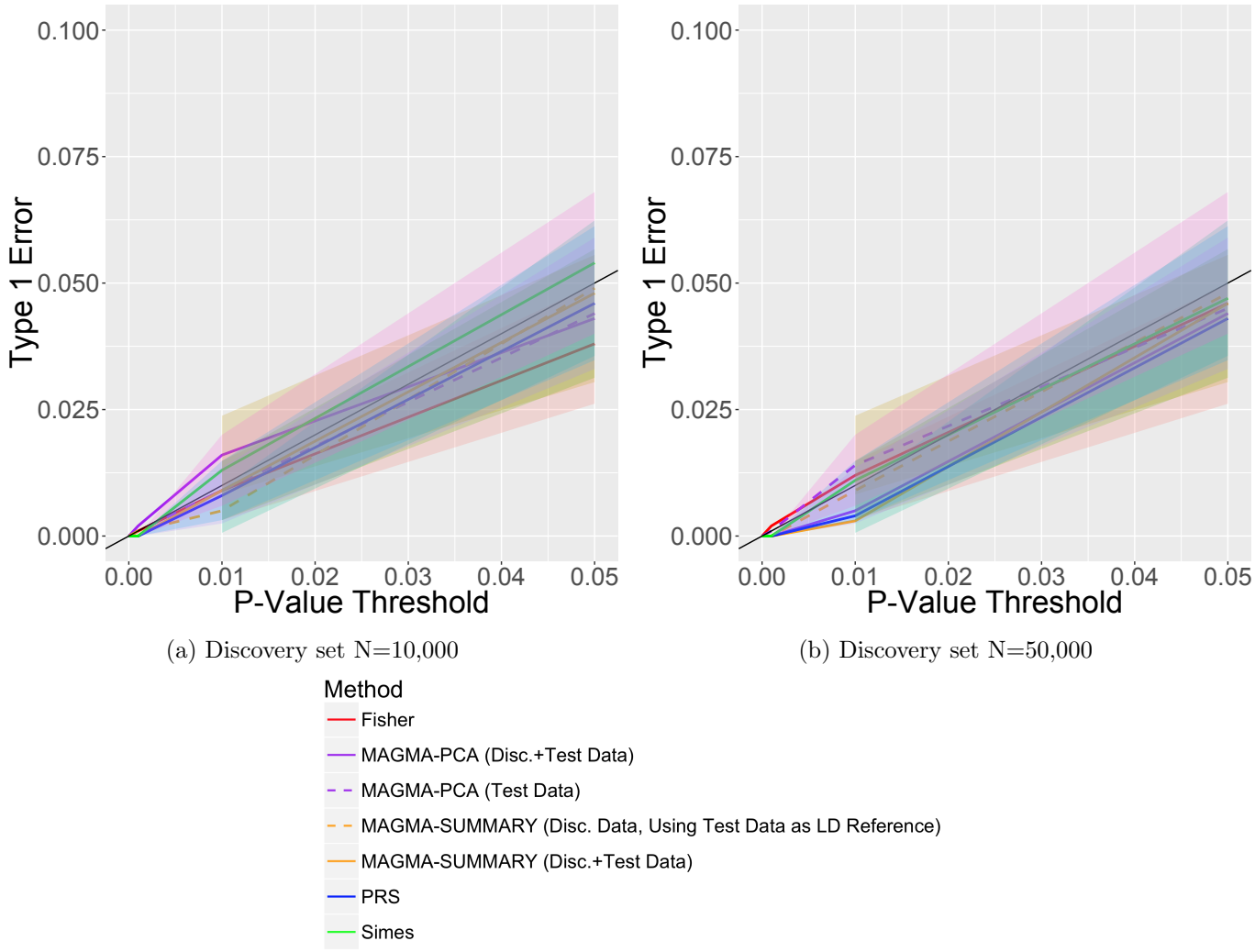
4.3 Results

4.3.1 Type I error

Initial simulations were used to determine the type I error for all set-based methods in all possible scenarios to ensure that the methods are comparable in terms of power. The expected type I error for a p-value threshold of 0.05 being approximately 5%, the type I error is deemed as reasonable if the nominal value falls within the 95% Confidence Interval (CI). The PRS set-based method is shown by a blue line, Fisher's method is displayed as a red line, Simes method is shown as a green line, MAGMA-PCA is displayed as a purple line, the solid purple line is in the combined test and discovery data and the dashed purple line is in the test set only, MAGMA-SUMMARY is shown as an orange line, again the solid orange line is in the combined test and discovery data and the dashed orange line is in the discovery data only, using the test data to estimate LD.

4.3.1.1 100 SNP Simulation

The first type I error simulation considers the case where 100 independent SNPs belong to the set. The results for this simulation are seen in Figure 4.6. It can be seen that the CIs for all methods contains 0.05. Therefore, all methods are comparable in terms of power.



(a) Discovery set N=10,000

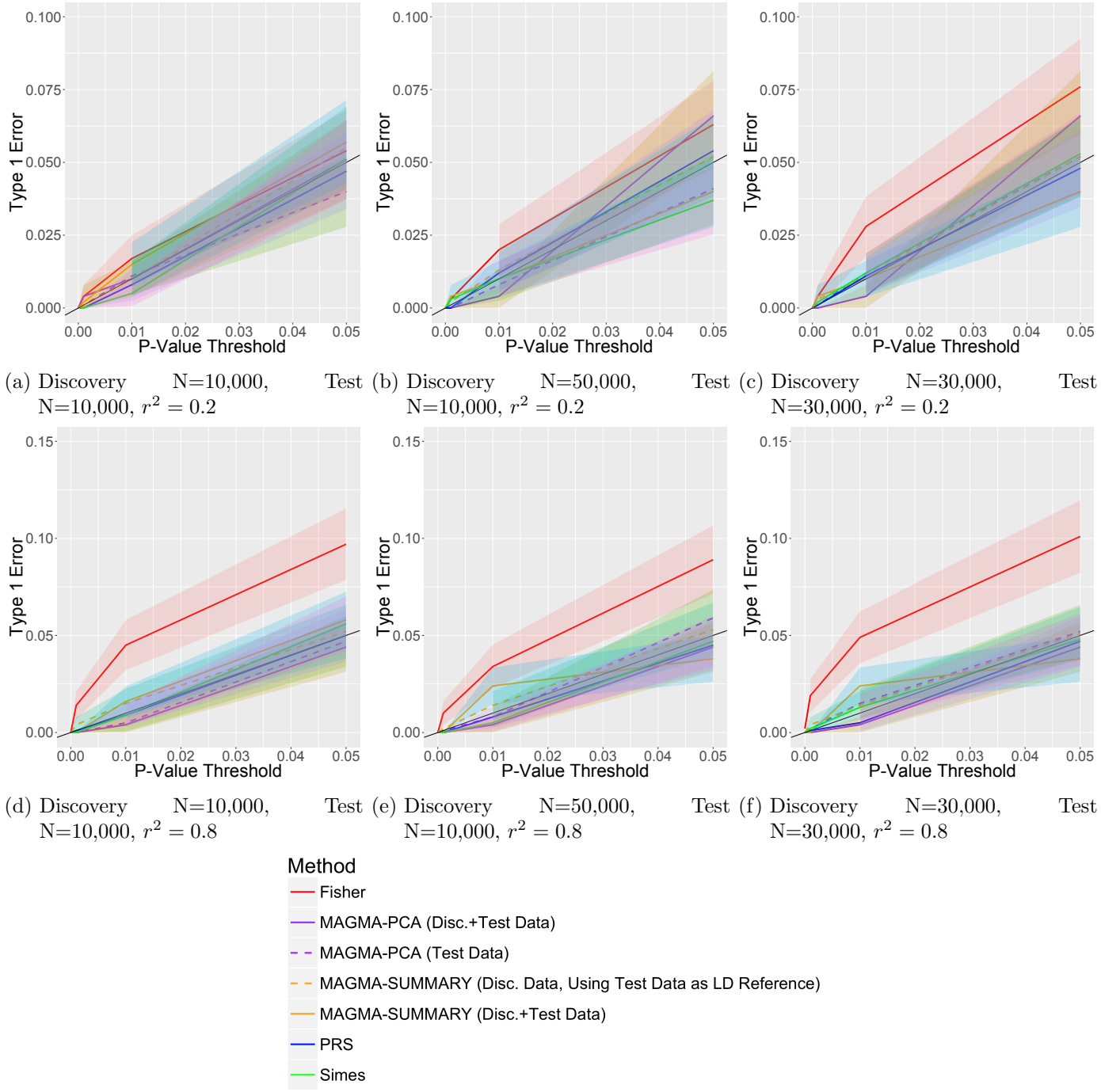
(b) Discovery set N=50,000

Figure 4.6: Type I Error Comparison of Set-Based Methods; Simulation of 100 independent SNPs where none are associated with disease with OR=1 and Test N=10,000. *Note: y-axis scale is not between 0 and 1*

4.3.1.2 Simple LD Block

The second type I error scenario represents an LD block. 100 SNPs were simulated which were not associated with disease, 10 SNPs are in LD and 90 SNPs are independent, where $r^2 = 0.2$ or $r^2 = 0.8$. Results can be seen in Figure 4.7. As expected, it is seen that the type I error is largely inflated for Fisher's method, this is due to the assumption of independence between SNPs. It is interesting that when LD between SNPs is $r^2 = 0.2$, this inflation is still seen. This LD strength was selected as it is often the threshold used for LD pruning, suggesting that Fisher's will be inflated, even after LD pruning the data prior to analysis. The inflation for Fisher's method increases with increasing LD

between SNPs. Type I error is elevated for the combined MAGMA-PCA method when the discovery sample size is 50,000 and $r^2 = 0.2$. The power estimates for Fisher's method will not be comparable to other methods due to this inflation, although, all other methods will be comparable to one another.



(a) Discovery N=10,000, Test N=10,000, $r^2 = 0.2$

(b) Discovery N=50,000, Test N=10,000, $r^2 = 0.2$

(c) Discovery N=30,000, Test N=30,000, $r^2 = 0.2$

(d) Discovery N=10,000, Test N=10,000, $r^2 = 0.8$

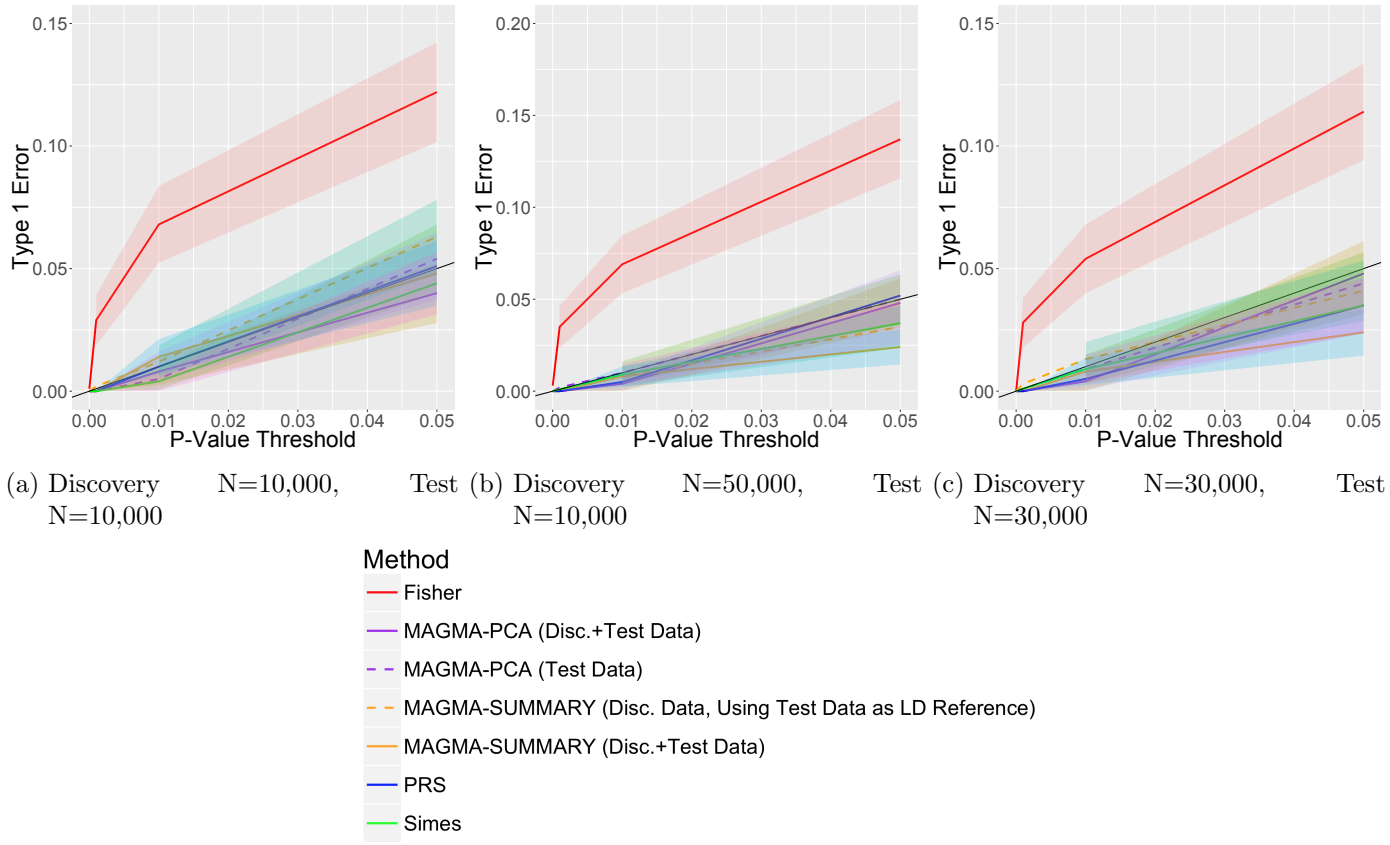
(e) Discovery N=50,000, Test N=10,000, $r^2 = 0.8$

(f) Discovery N=30,000, Test N=30,000, $r^2 = 0.8$

Figure 4.7: Type I Error Comparison of Set-Based Methods; Simulation of 10 SNPs in LD and 90 independent SNPs. Figures 4.7a, 4.7b and 4.7c have $r^2 = 0.2$, and Figures 4.7d, 4.7e and 4.7f have $r^2 = 0.8$. Figures 4.7a and 4.7d have a discovery and test sample size of 10,000, Figures 4.7b and 4.7e have a discovery set N=50,000 and test set N=10,000 and Figures 4.7c and 4.7f have discovery and test sets with N=30,000. *Note: y-axis scale is not between 0 and 1*

4.3.1.3 Complex LD Structure

The next scenario represents a complex LD structure. There are a total of 100 SNPs; 10 SNPs in LD with $r^2 = 0.2$, 10 SNPs in LD with $r^2 = 0.4$, 10 SNPs in LD with $r^2 = 0.6$ and 10 SNPs in LD with $r^2 = 0.8$, and 60 independent SNPs. The type I error graph is seen in Figure 4.8. The type I error is clearly highly inflated for Fisher's method, so this is not comparable with other methods.



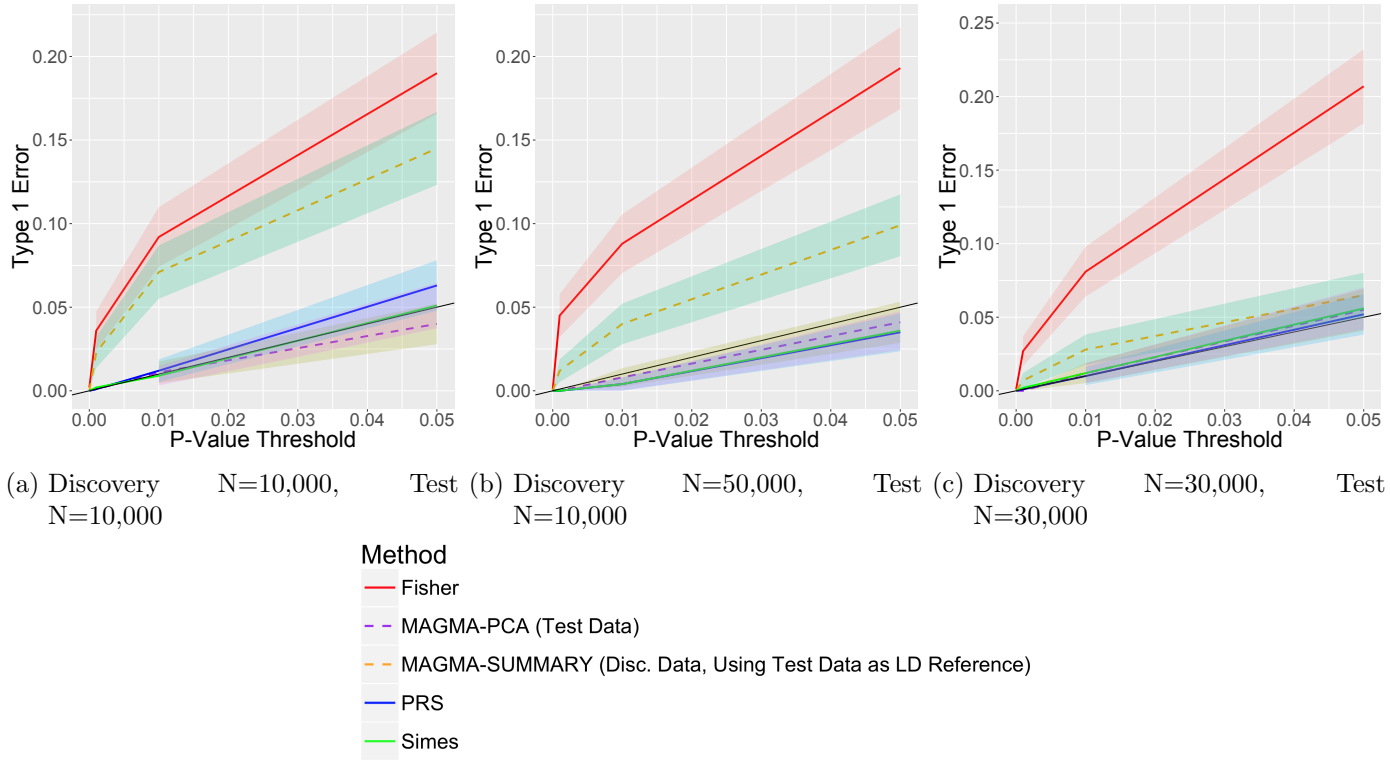
(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 4.8: Type I Error Comparison of Set-Based Methods; Simulation of 10 SNPs in LD with $r^2 = 0.2$, 10 SNPs in LD with $r^2 = 0.4$, 10 SNPs in LD with $r^2 = 0.6$, 10 SNPs in LD with $r^2 = 0.8$ and 60 independent, unassociated SNPs. *Note: y-axis scale is not between 0 and 1.*

4.3.1.4 Different LD structure of Discovery and Test Datasets

This next scenario is to represent discovery and test sets which come from differing populations. Therefore, the LD structure differs between the discovery set ($r^2 = 0.8$) and the test set ($r^2 = 0.6$). Type I error can be seen in Figure 4.9. Here, Fisher's method shows

the largest inflation, and MAGMA-SUMMARY approach also shows some inflation when the method is used in the discovery data using the test set to estimate LD. This is likely due to the MAGMA-SUMMARY set adjusting the discovery data with LD estimates from the test set (which has lower LD) and therefore, the correction is not large enough.

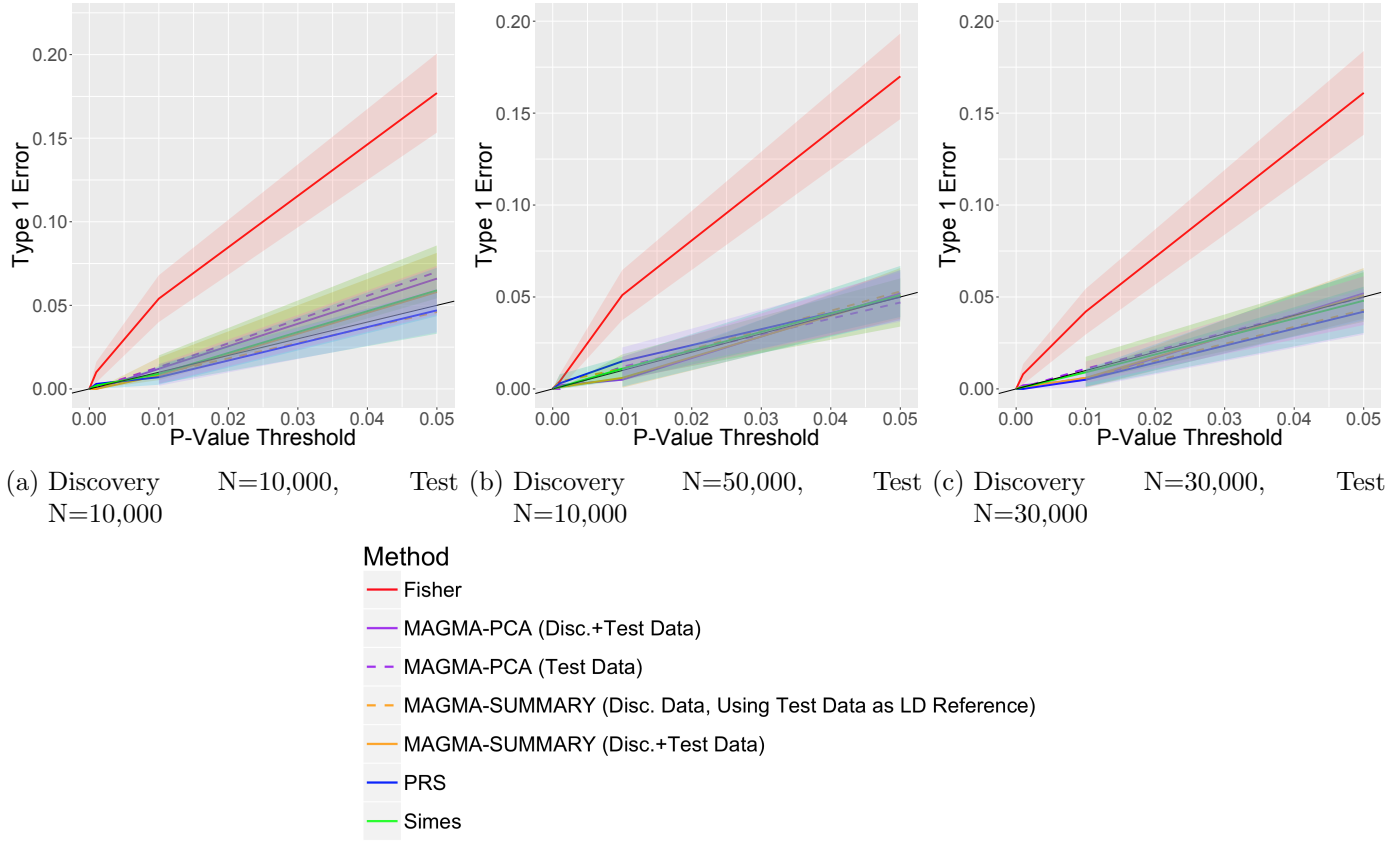


(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 4.9: Type I Error Comparison of Set-Based Methods; Simulation of 10 SNPs in LD with OR$\sim$ N(1, 0.2^2) and 90 independent, unassociated SNPs where Test LD is moderate ($r^2 = 0.6$) and Discovery LD is high ($r^2 = 0.8$). *Note: y-axis scale is not between 0 and 1.*

4.3.1.5 Effect Sizes with Varying Direction

This situation considers where effect sizes are not necessarily in the same direction due to varying MAF [88]. Set-based methods which only use p-values are unable to take the direction of the SNP effect into account. Therefore, the 10 SNPs which are in LD have an association which randomly varies in direction. The type I error for this simulation is seen in Figure 4.10. Again, it is clear that Fisher's method has inflated type I error in this case.



(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 4.10: Type I Error Comparison of Set-Based Methods; Simulation of 10 SNPs with Varying LD with ORs with Randomly Varying Direction and 90 independent, unassociated SNPs. *Note: y-axis scale is not between 0 and 1*

4.3.1.6 Real Data Simulation

The final scenario uses real data to assess a true LD structure. 129 SNPs are taken from real AD GERAD data, and the case-control status is permuted in order to remove the effect size of any SNPs. The LD structure of these 129 SNPs is seen in Figure 4.5. These 129 SNPs are an area of strong LD (chr1, 50,002,165:52,034,812), full details of these 129 SNPs can be seen in Supplementary Table 11.1. The type I error is seen in Figure 4.11. It is clear that Fisher's method has largely inflated type I error, and all other methods are reasonable.

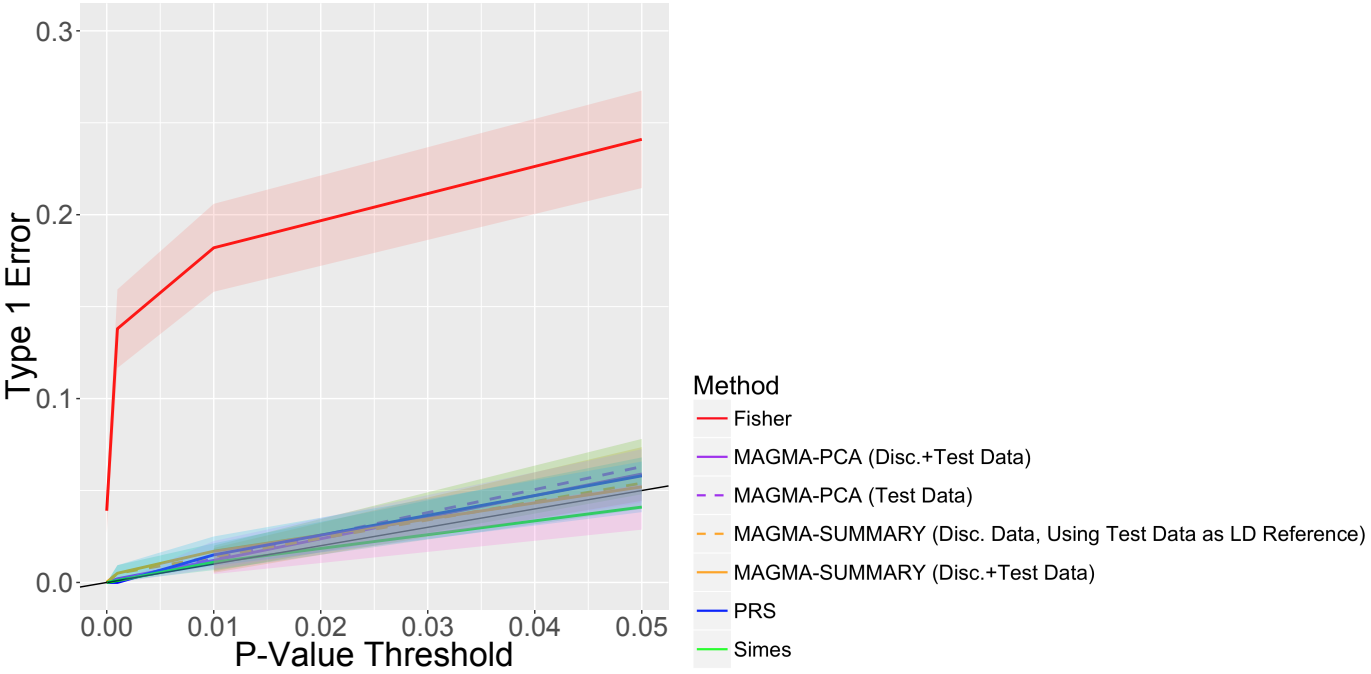


Figure 4.11: Type I Error Comparison of Set-Based Methods; Simulation 129 SNPs from Real Data, with Permuted Phenotypes to Remove Effect Sizes, N=13,164. *Note: y-axis scale is not between 0 and 1*

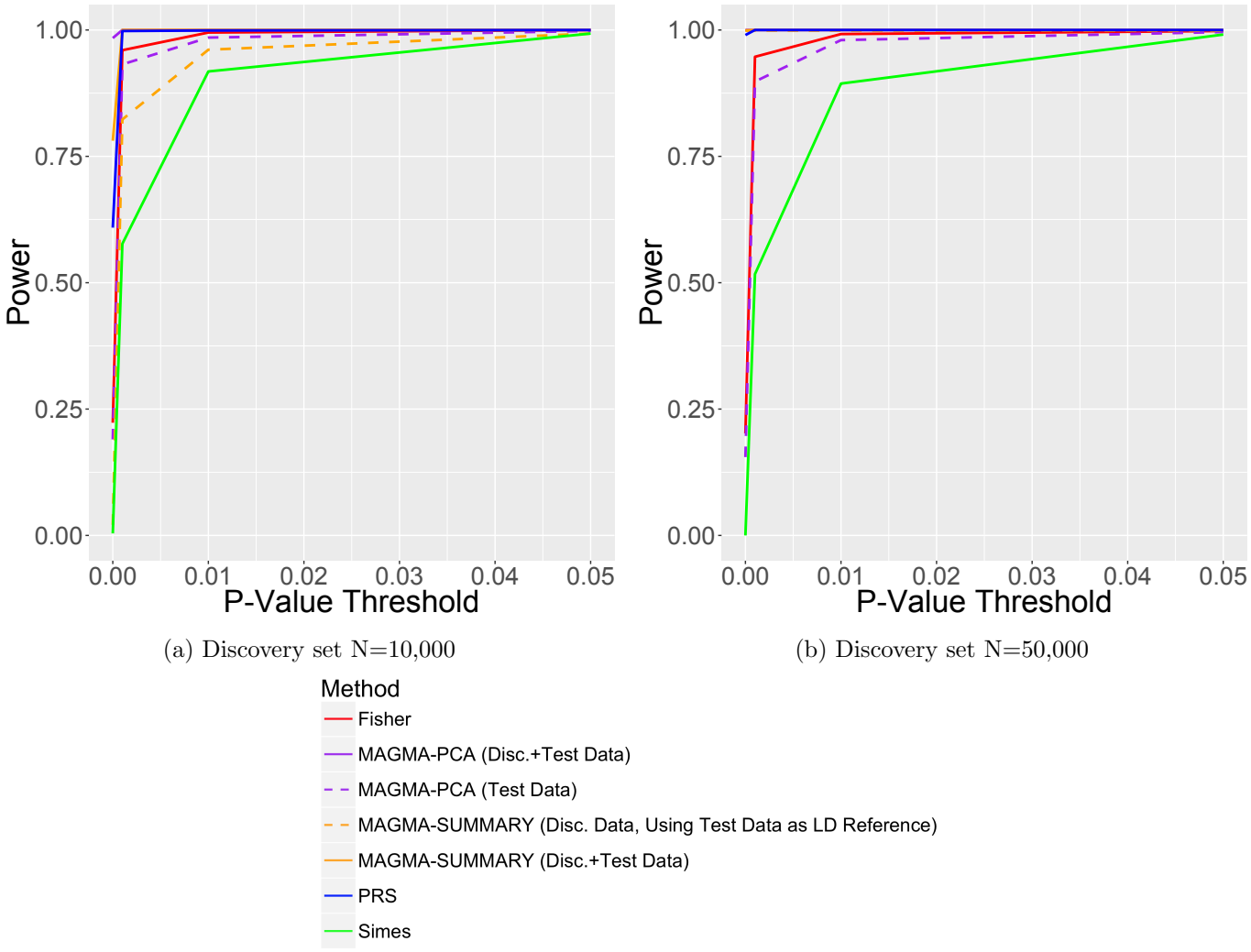
4.3.2 Power Comparison

The power of the $\text{PRS}_{set-based}$ method with MAGMA-PCA in the combined set and in the test set only, MAGMA-SUMMARY in the combined set, MAGMA-SUMMARY in the discovery set, estimating LD from the test set, and Fisher and Simes methods in the test set only, is compared using the same simulated scenarios as before, but where some SNPs have an association with disease.

4.3.2.1 100 SNP Simulation

Figure 4.12 shows the simulation of 100 independent SNPs where 10 of these SNPs are associated with disease with OR=1.1. All methods have relatively high power in this case, since the effect of the 10 SNPs is greater than the effect of the noise of the remaining 90 SNPs. The $\text{PRS}_{set-based}$ method (solid blue line) has higher or equivalent power compared to all other methods, MAGMA-PCA in the combined data (solid purple line) also has very high power, since this is also informed by the discovery set. Simes (solid green line) has

the lowest power compared to the other methods, this is likely because Simes corrects for the number of SNPs in the set.



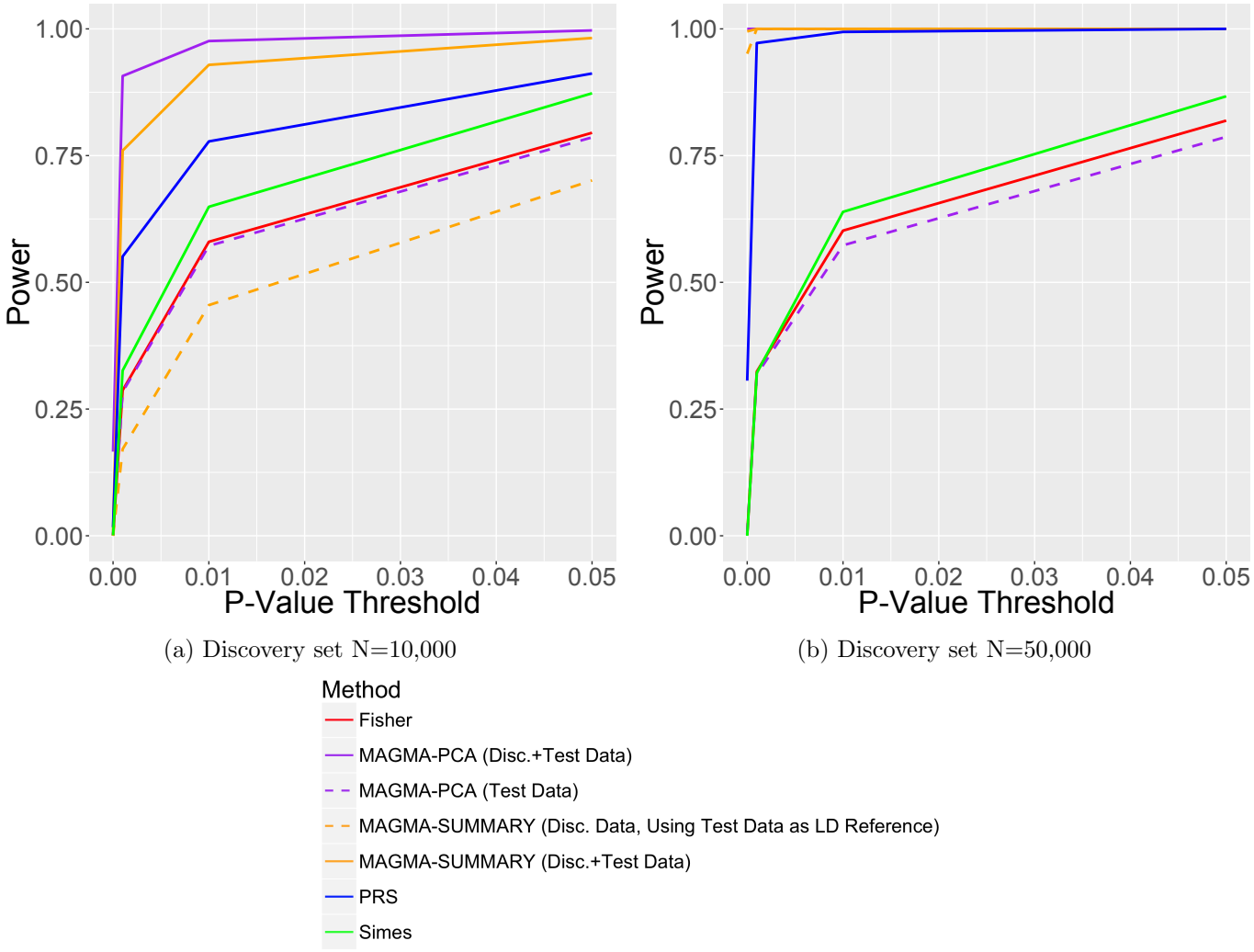
(a) Discovery set N=10,000

(b) Discovery set N=50,000

Figure 4.12: Power Comparison of Set-Based Methods; Simulation of 100 independent SNPs where 10% are associated with disease with OR=1.1 and Test N=10,000.

The next case, seen in Figure 4.13, considers when only 5% of all independent SNPs within the set are associated with disease, with an OR=1.1. The set-based methods which make use of the discovery set have higher power than those calculated in the test set only. The $PRS_{set-based}$ method has equivalent power compared to other set-based methods when the discovery set has N=50,000. The combined MAGMA-PCA method has higher power than the $PRS_{set-based}$ method when the discovery sample size is 10,000. Simes has higher power than Fisher's (solid red line) method and MAGMA-SUMMARY method (dotted purple line), this opposes the result seen previously. This is since the methods are testing

different hypotheses, the alternative hypothesis for Simes is that **at least** one SNP within the set is associated with disease, whereas Fisher's alternative hypothesis is that the set is associated with disease, therefore Simes will be unaffected by noise from SNPs not associated with disease.



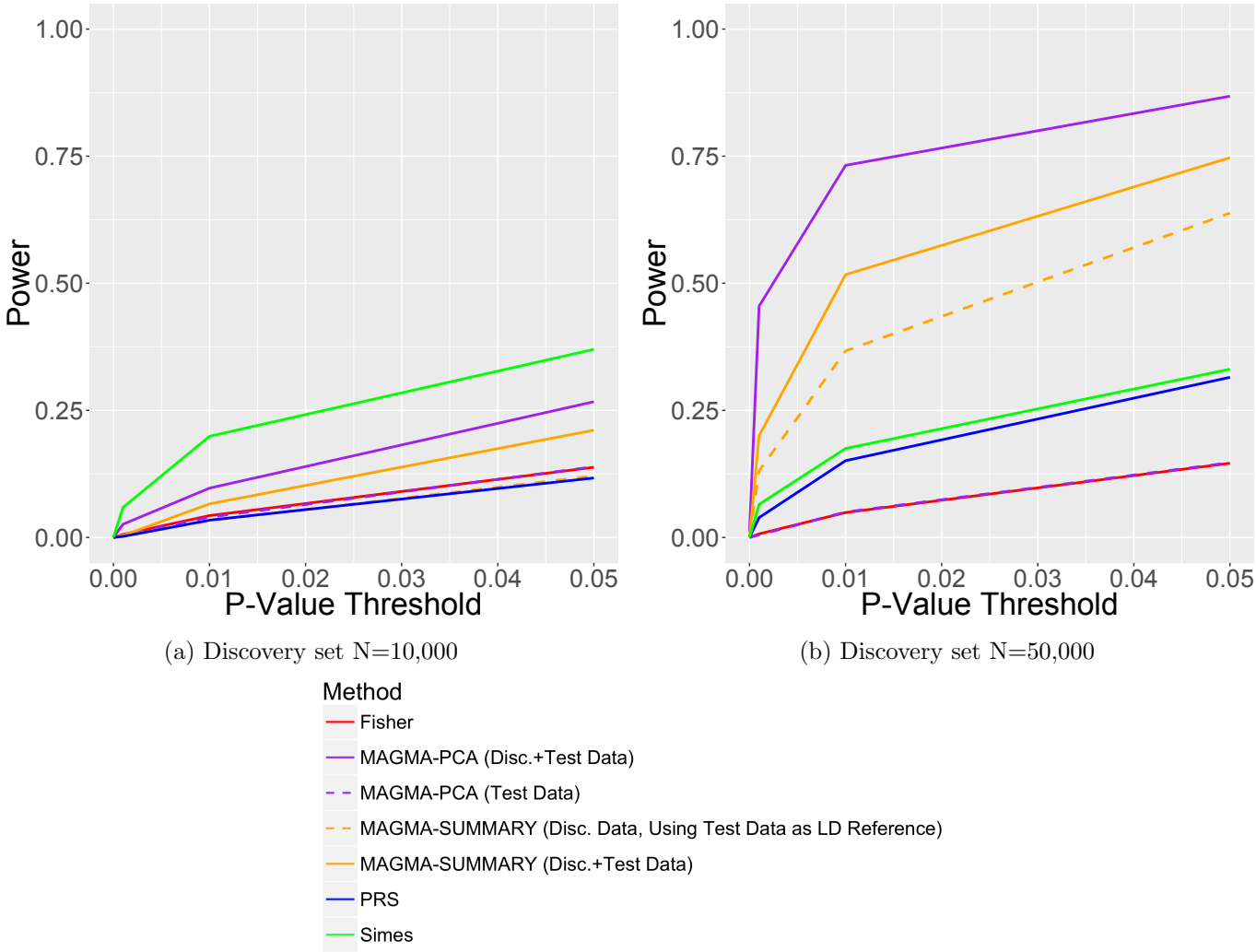
(a) Discovery set N=10,000

(b) Discovery set N=50,000

Figure 4.13: Power Comparison of Set-Based Methods; Simulation of 100 independent SNPs where 5% are associated with disease with OR=1.1 and Test N=10,000.

In the final case, shown in Figure 4.14, there are 100 independent SNPs but only one is associated with disease. Here, most methods have much lower power, due to the noise of the large number of unassociated SNPs. Simes has relatively good power, again, because this method is testing whether at least one SNP is associated with disease, so is less affected by noise. When the discovery set is larger (N=50,000) the MAGMA-PCA method in the combined set has the highest power, this is likely because it utilises all available raw

genotype data. MAGMA-SUMMARY also has good power in this case, likely because it is using the discovery set summary statistics which are based on a larger set than the methods computed in the test set.



(a) Discovery set N=10,000

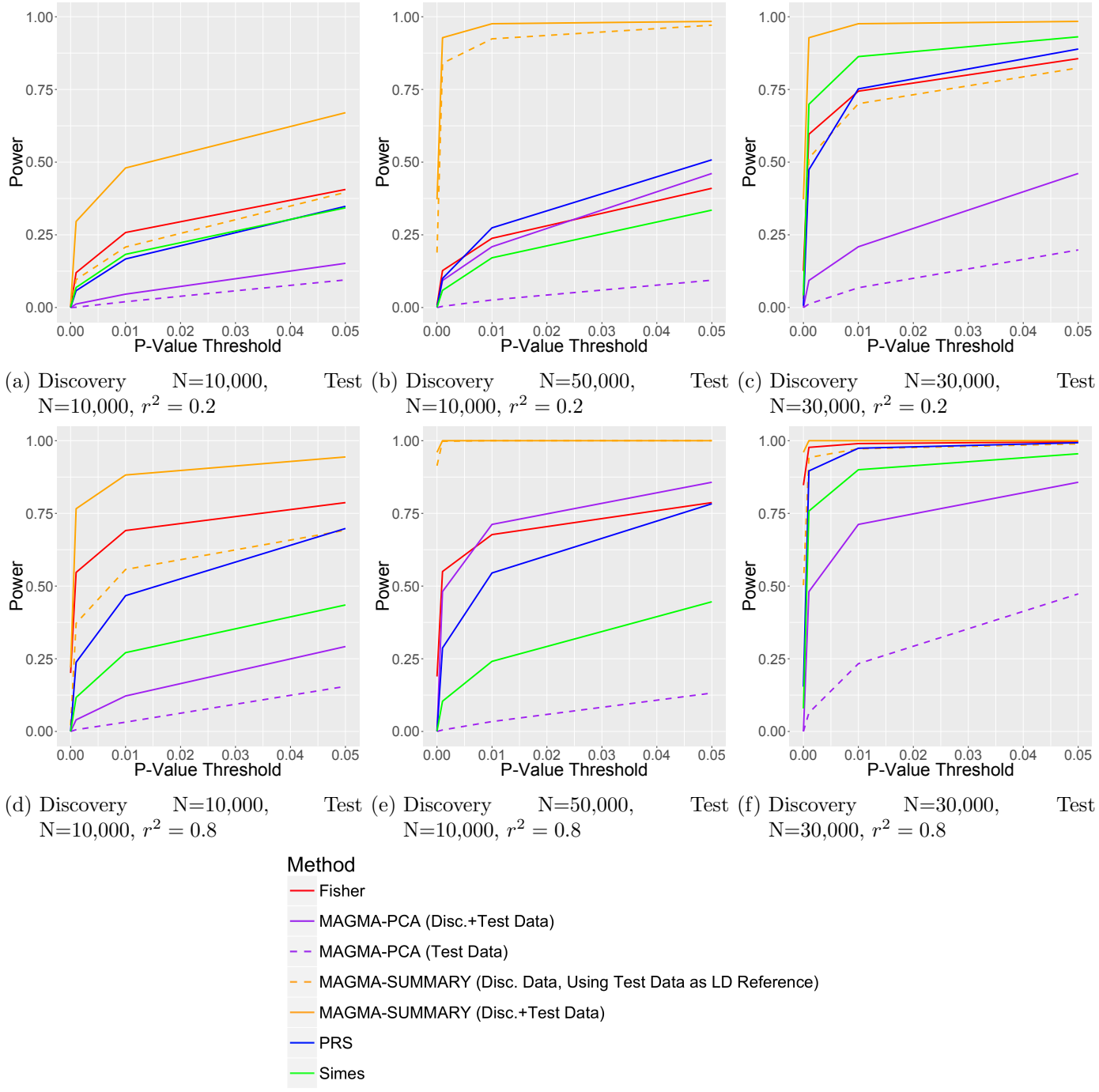
(b) Discovery set N=50,000

Figure 4.14: Power Comparison of Set-Based Methods; Simulation of 100 independent SNPs where 1% are associated with disease with OR=1.1 and Test N=10,000.

4.3.2.2 Simple LD Block

Figure 4.15 shows the case where 10 SNPs were simulated in LD, with $r^2 = 0.2$ and OR=1.1, and 90 independent SNPs which have no association with disease. A relatively low value of $r^2 = 0.2$ was chosen since this is a commonly used threshold for LD pruning in practice for genotype data. As the size of the discovery set increases, the power for the $\text{PRS}_{set-based}$ method increases and is higher than all methods calculated in the test

set only. Fisher's method has quite high power, but this should be disregarded due to the inflated type I error. MAGMA-SUMMARY in the combined test and discovery sets has highest power in all cases; it is odd how much higher the power is compared to the MAGMA-PCA approach in the same data. MAGMA-SUMMARY in the discovery set only has particularly high power compared to other methods when the test set N=10,000 and the discovery set N=50,000, this is because this approach is computing set effects in the discovery set, which has a larger sample size relative to the test set.



(a) Discovery N=10,000, Test N=10,000, $r^2 = 0.2$

(b) Discovery N=50,000, Test N=10,000, $r^2 = 0.2$

(c) Discovery N=30,000, Test N=30,000, $r^2 = 0.2$

(d) Discovery N=10,000, Test N=10,000, $r^2 = 0.8$

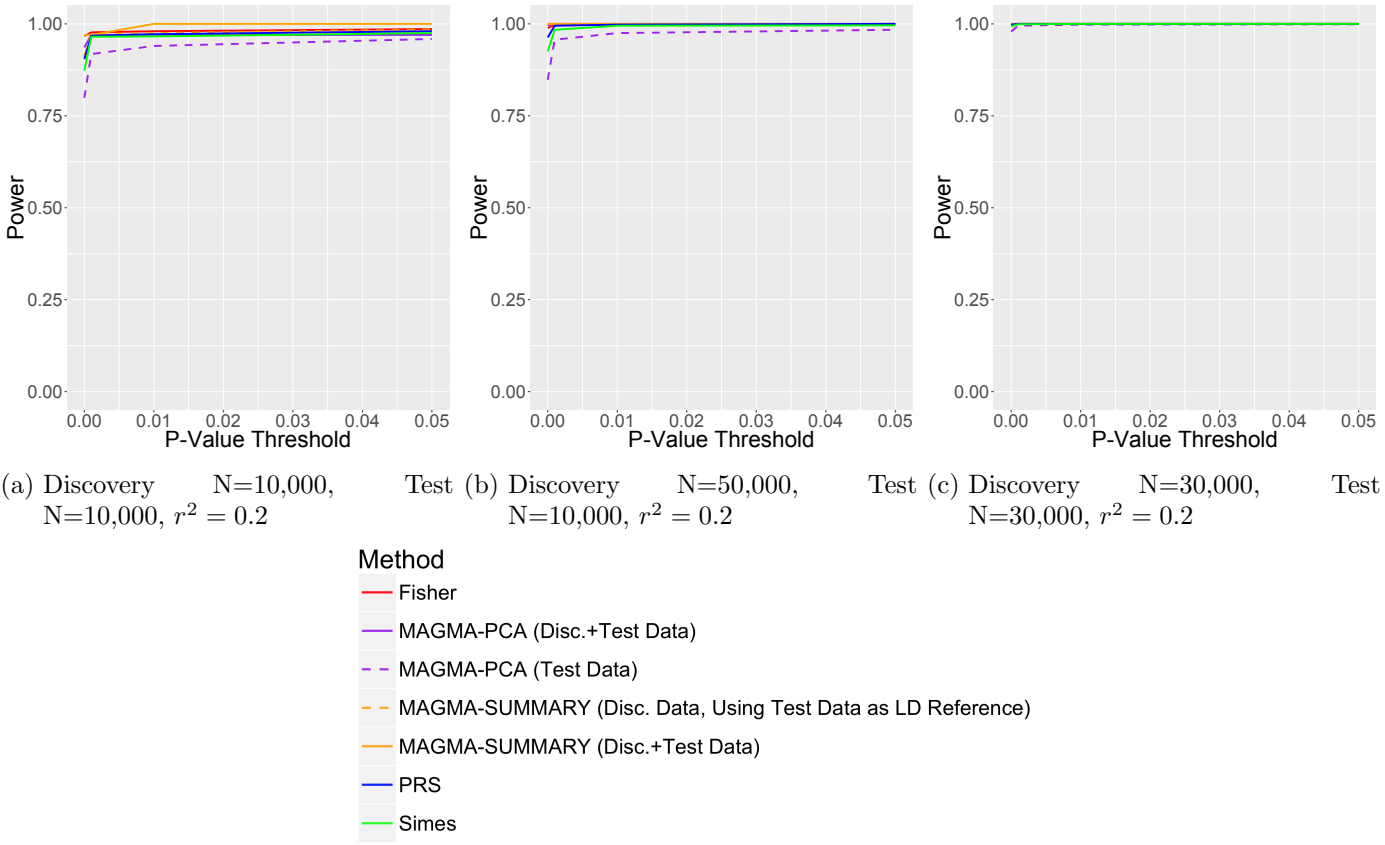
(e) Discovery N=50,000, Test N=10,000, $r^2 = 0.8$

(f) Discovery N=30,000, Test N=30,000, $r^2 = 0.8$

Figure 4.15: Power Comparison of Set-Based Methods; Simulation of 10 SNPs in LD and 90 independent SNPs. Figures 4.15a, 4.15b and 4.15c show Simple LD Structure Simulations where $r^2 = 0.2$, and Figures 4.15d, 4.15e and 4.15f show Simple LD Structure Simulation where $r^2 = 0.8$. Figures 4.15a and 4.15d have a discovery and test sample size of 10,000, Figures 4.15b and 4.15e have a discovery set N=50,000 and test set N=10,000 and Figures 4.15c and 4.15f have discovery and test sets with N=30,000.

4.3.2.3 Complex LD Structure

The next scenario represents a complex LD structure. There are a total of 100 SNPs; 10 SNPs in LD with $r^2 = 0.2$, 10 SNPs in LD with $r^2 = 0.4$, 10 SNPs in LD with $r^2 = 0.6$ and 10 SNPs in LD with $r^2 = 0.8$, and 60 independent SNPs. The 40 SNPs which are in LD blocks are all associated with disease, with OR$\sim$ N(1.02, 0.2^2), and the remaining 60 independent SNPs are unassociated with disease. The power graph is seen in Figure 4.16. The power is high for all set-based methods, and it is therefore difficult to distinguish between the approaches. When the test set is smaller (N=10,000) the MAGMA-PCA in the test set only has the lowest power.



(a) Discovery N=10,000, Test N=10,000, $r^2 = 0.2$

(b) Discovery N=50,000, Test N=10,000, $r^2 = 0.2$

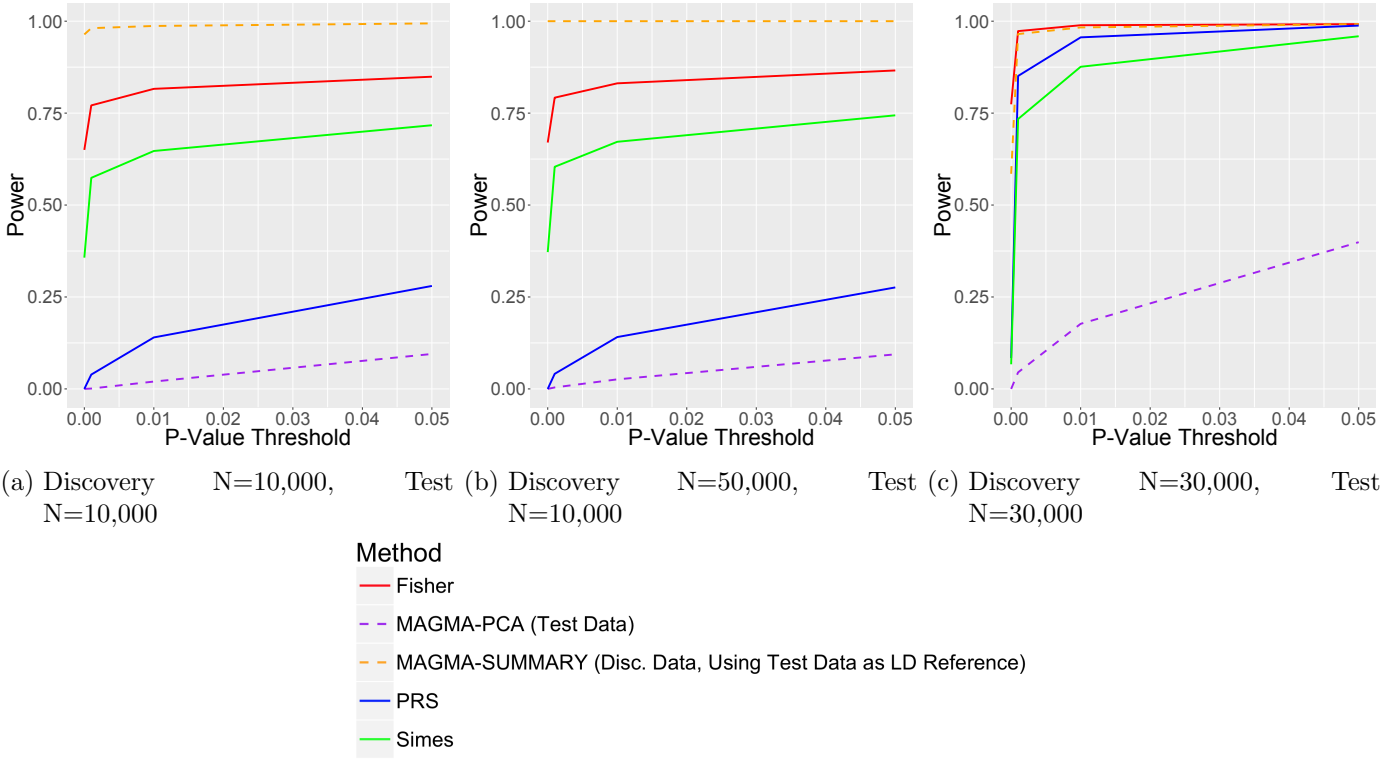
(c) Discovery N=30,000, Test N=30,000, $r^2 = 0.2$

Figure 4.16: Power Comparison of Set-Based Methods; Simulation of 10 SNPs in LD with $r^2 = 0.2$, 10 SNPs in LD with $r^2 = 0.4$, 10 SNPs in LD with $r^2 = 0.6$, 10 SNPs in LD with $r^2 = 0.8$, all 40 SNPs in LD Blocks have OR$\sim N(1.02, 0.36)$, and 60 independent, unassociated SNPs.

4.3.2.4 Different LD Structure of Discovery and Test Datasets

In practice it is unlikely that the discovery and test datasets will be obtained from identical populations. Thus, simulations are undertaken to determine the influence on power for each of the set-based methods. Again, 10 SNPs in LD which are associated with disease and 90 independent unassociated SNPs are simulated. The 10 SNPs which are associated with disease have an OR which comes from a Normal distribution with mean=1.1 and SD= 0.2, so the ORs for individual SNPs differ across populations. The LD in the discovery set is high with $r^2 = 0.8$ and is moderate in the test set with $r^2 = 0.6$. The MAGMA-PCA and MAGMA-SUMMARY methods in the combined test and discovery sets are not presented here, since the combination would result in one set with intermediate LD which would not be comparable.

These results are seen in Figure 4.17. The $\text{PRS}_{set-based}$ method has higher power compared to MAGMA-PCA in the test set only and has higher power compared to Simes when the test and discovery set have a sample size of 30,000. The MAGMA-SUMMARY method has highest power but this is due to the inflated type I error. Fisher's has equivalent power to the MAGMA-SUMMARY method when the test and discovery set have N=30,000. Simes is the method with the highest power which also has reasonable type I error. The PRS method has an increase in power when the test set is larger (N=30,000 compared to N=10,000), the test set is used to estimate LD and also adds power to the logistic regression model which is why this is seen.



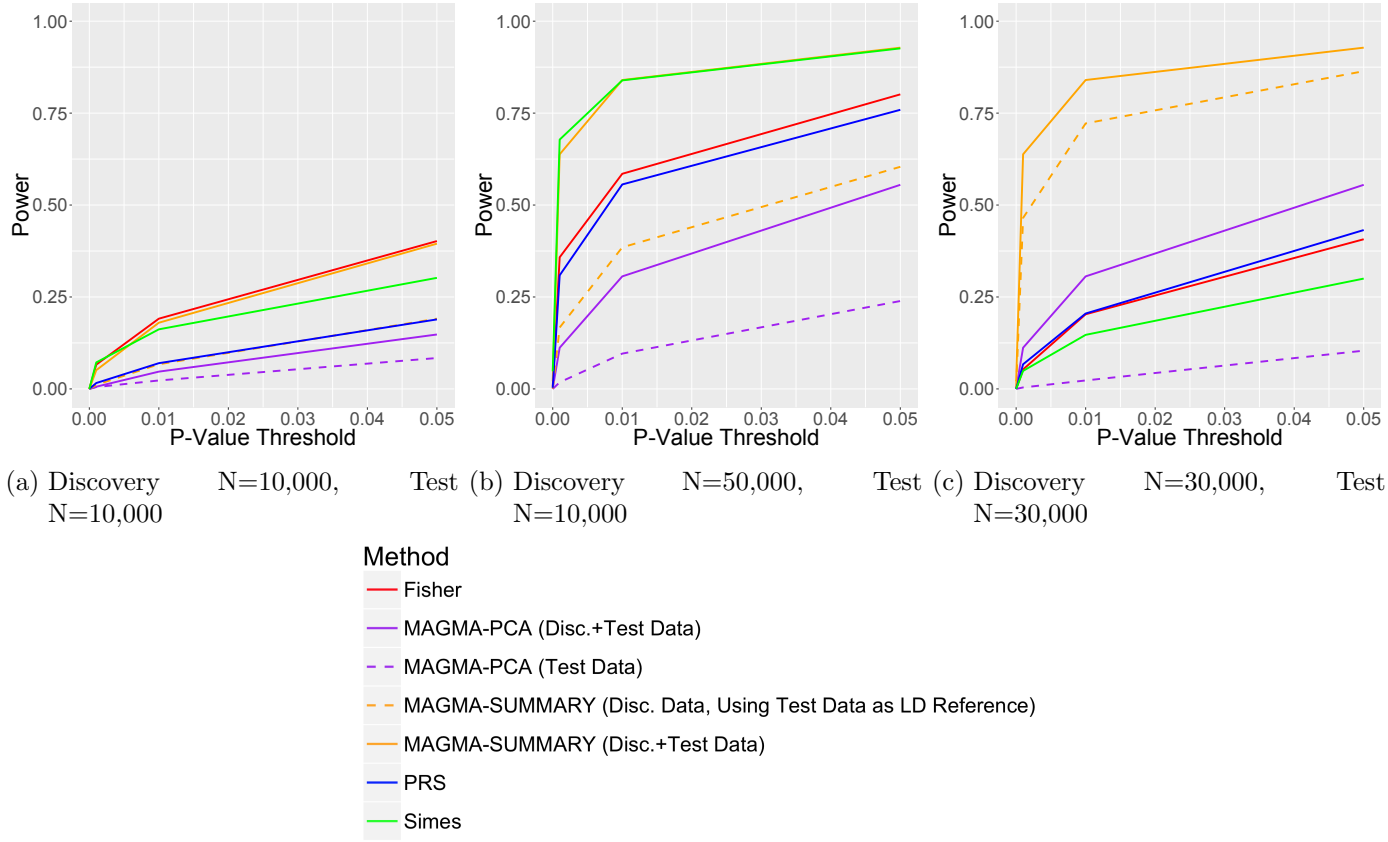
(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 4.17: Power Comparison of Set-Based Methods; Simulation of 10 SNPs in LD with OR$\sim$ N(1.1, 0.2^2) and 90 independent, unassociated SNPs where Test LD is moderate ($r^2 = 0.6$) and Discovery LD is high ($r^2 = 0.8$).

4.3.2.5 Effect Sizes with Varying Direction

Since real data have SNPs which have varying directions (both risk and protective SNPs), the next simulation represents this case. As before, 10 SNPs are simulated which have randomly varying effect sizes and have differing strengths of LD. The 10 SNPs in LD are associated with disease. The results can be seen in Figure 4.18.

Clearly, the methods which only utilise the p-values have very high power. This is because the direction of effect is not incorporated into the analysis, and therefore effects which should cancel are being overemphasised. Both the $\text{PRS}_{set-based}$ method and MAGMA-PCA method incorporate the direction of effect into the set-based analysis, of these two methods, the $\text{PRS}_{set-based}$ method has higher power compared to the MAGMA-PCA method in the test set only. The $\text{PRS}_{set-based}$ method has higher power than MAGMA-PCA in the combined test and discovery sets, when the test sample size is 10,000.



(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 4.18: Power Comparison of Set-Based Methods; Simulation of 10 SNPs with Varying LD with ORs with Randomly Varying Direction and 90 independent, unassociated SNPs.

4.3.2.6 Real Data Simulation

The final scenario is created from real data, in order to investigate simulations with a true LD structure. 129 SNPs from the AD GERAD data were used. The 10 most associated SNPs in both directions are used to maintain the effect size. A percentage of cases and controls had their genotype maintained, and the remainder were permuted; this enabled an effect size to remain while generating new datasets. The results are seen in Figure 4.19.

From the real data simulations, it is seen that the $PRS_{set-based}$ method has the highest power of all the set-based methods when a true LD structure is simulated. The power for both $PRS_{set-based}$ methods, MAGMA-PCA and MAGMA-SUMMARY in combined data are much higher than other set-based methods.

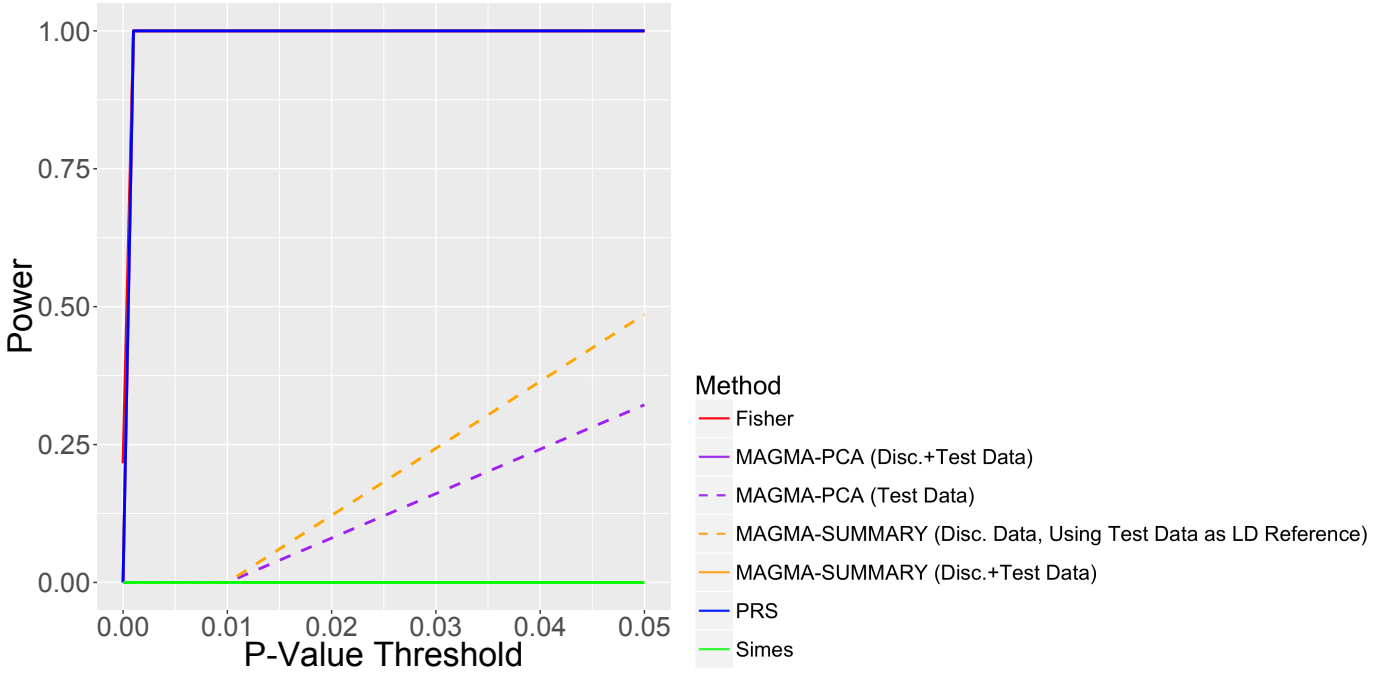


Figure 4.19: Power Comparison of Set-Based Methods; Simulation 129 SNPs from Real Data, with Permuted Phenotypes to Maintain Effect Sizes, N=13,164.

4.4 Discussion

This chapter focuses on the use of PRS as a set-based approach. This produces a risk score per person per set, and informs the set with additional discovery data whilst maintaining a self-contained test in the test data.

The PRS method requires SNPs to be independent, and therefore, SNPs should be pruned for LD. Although, this is the choice of the researcher. The simulations presented here with LD between SNPs of $r^2 = 0.2$ represents the 'pruned' case, since 0.2 is a regularly used r^2 threshold used in pruning. The simulations presented with LD of $r^2 = 0.8$ represent the 'unpruned' case. The type I error for the PRS method in both these cases is similar, but the power is higher in the unpruned case. When considering the simulations with the real LD structure, the type I error for the PRS method is slightly inflated.

The $\text{PRS}_{set-based}$ method has reasonable type I error in all simulations considered, although it is well known that results can be inflated in the presence of LD [72]. The difference here may be to do with the size of the set or the strength of LD. Therefore, it

would still be advisable to LD prune before using this $PRS_{set-based}$ method.

The $PRS_{set-based}$ method has relatively good power compared to all other set-based methods investigated here. Most importantly, it improves power upon the MAGMA-PCA in the test data only, so this method is an improvement over the MAGMA approach investigated in Chapter 3. Additionally, in the simulations with the real data LD structure, the $PRS_{set-based}$ method has very high power. Therefore, it may be possible to determine genes or pathways which are associated with AD when using real data. The $PRS_{set-based}$ method has increasing power as the discovery set sample size increases, this is due to the increased accuracy of effect size estimates in discovery set. This method has the potential to discover data-driven pathways, it may be used for prioritisation of subjects for follow-up studies (e.g. clinical trials) based upon single gene or pathway PRS, and for prioritising genes for further functional studies (e.g. animal models).

The set-based methods all use slightly different data, and therefore, they are difficult to compare. In a few cases the combined MAGMA-PCA method had higher power than the $PRS_{set-based}$ method, however, this requires raw genotype data for both the discovery and test set, which often will not be available. Fisher's method is inappropriate in the presense of LD since it assumes independence between SNPs and has shown very inflated type I error in the presense of LD. Simes is an appropriate method in the presense of LD between SNPs, since it considers only whether at least one SNP in the set is associated with disease. Simes generally has less power since it corrects for False Discovery Rate (FDR) (this is better than the degrees of freedom penalty in Fisher's method) in the presense of LD. The $PRS_{set-based}$ method has fewer false positive associations, since the test set is informed by the discovery set, false positives are cancelled out whereas this does not occur in Simes/Fisher's methods.

The $PRS_{set-based}$ method should be used in LD pruned data as is standard practice when using PRS, since inflated type I error was observed in the case of a real data LD structure [72][89][90]. Data can be pruned for LD in two different ways; the first is to randomly remove SNPs which are in LD above a particular threshold, and the second is intelligent pruning, which retains the SNPs which are most strongly associated with disease (using the

--clump option in PLINK). When a number of SNPs are all highly associated with disease, it is likely that there will be a correlation between these SNPs due to this association, rather than simply LD. It would therefore be beneficial to improve upon this method by adjusting for LD between SNPs, without removing the association of SNPs, thus removing the need to prune the data and increasing power by using a larger number of SNPs.

5 PRS Approach: Gene-Based and Pathway Analyses in AD Data

5.1 Introduction

Set-based analyses are a valuable alternative to single-SNP analyses, since the effects of all SNPs in the gene are combined. This has the potential to improve the power of the analysis since the overall effect of the set is larger than that of a single SNP. In addition, gene-based analysis, a gene centred equivalent of set-based analysis, identifies genes associated with disease rather than a single SNP as a proxy for the gene. In gene-based analyses, genes are found to be fairly consistently associated with disease across different populations. In contrast, different SNPs in a set in LD may be found to be associated with a disease in different samples. Gene-based analyses also directly provide information for functional analysis [44]. Set-based analysis can also be employed as a pathway analysis, and applied to sets of SNPs defined by epigenomics for different tissue/cell types.

Gene-based analysis using Brown's method [52] have been applied to the imputed IGAP stage 1 and 2 data [21]. Brown's method combines the p-values for SNPs in a gene whilst adjusting for LD between SNPs. This analysis found additional genes associated with AD compared to those from the single-SNP IGAP analysis [20].

A pathway analysis has also been previously published in the AD IGAP data. This used the ALIGATOR approach [61]; this determines which genes contain at least one SNP below a particular p-value threshold and compares these to a random gene set to determine whether the gene set of interest is enriched in AD. Eight pathways were determined from this

analysis; these were the immune response, regulation of endocytosis, cholesterol transport, hematopoietic cell lineage, proteasome-ubiquitin activity, reactome hemostasis, clathrin and protein folding pathways [23][24].

It is possible that the basic principles of polygenic score analysis can also be applied to individual genes, or to gene-set analyses. The motivation for doing so is somewhat different than PRS analyses of genome-wide data; rather than predict case-control status or trait liability captured, the goal in applying the principles of PRS to genes and gene sets is to detect association to these potentially biologically informative features. As for genome-wide analyses, genes or gene-set PRS could be used to predict affected status, or to estimate the gene-specific or pathway-specific SNP liability captured by GWAS. However, for polygenic disorders where risk is dispersed across hundreds of genes and multiple gene-sets, self-evidently, gene-specific or pathway-specific SNP liability will be lower than the liability captured by genome-wide data, and accordingly, such tests will afford less case-control discriminatory power. Risk scores for each individual per set can be used to stratify individuals for follow-up studies and prioritise genes for further functional studies.

The approach of using PRS as a set-based method has been shown to be more powerful compared to MAGMA-PCA, Fisher and Simes methods in data with a real LD structure in Chapter 4. The $\text{PRS}_{set-based}$ method is able to determine the association of a set with disease whilst also producing a risk score per subject per set. By producing a risk score, it is possible to find commonality between genes or pathways using a correlation.

Set-based analyses can be implemented in either directly genotyped or imputed data. Imputed data has the main advantage that the number of SNPs can be vastly increased with minimal costs. Therefore, the power of the analyses can be increased with the additional SNPs. The issues with imputation are that it cannot be done with perfect accuracy, so it is possible to introduce error into the analyses, the impact of this can be reduced by quality checking the data and removing any SNPs which are imputed with too low quality. Additionally, increasing the number of SNPs is a benefit in terms of power, but may introduce bias as some set-based methods have inflated p-values simply due to

the number of SNPs in the set.

5.1.1 Objectives

This Chapter focuses on using the $PRS_{set-based}$ approach in real AD data by aiming to:

- Perform a gene-based analysis using the $PRS_{set-based}$ approach in the GERAD genotype data.
- Compare $PRS_{set-based}$ real data gene-based results with MAGMA-PCA, MAGMA-SUMMARY, Simes and Fisher's methods.
- Perform gene-based analysis in the GERAD imputed data.
- Compare the gene-based results using genotyped and imputed data.
- Investigate whether the $PRS_{set-based}$ method is biased by the number of SNPs in the gene and compare this to other set-based methods.
- Demonstrate whether genes from the gene-based analysis are represented in conserved regions.
- Compute PRS pathway scores for the eight pathways previously found to be associated with AD [23][24] and determine whether the PRS pathways are also associated with AD.
- Determine whether any correlation is observed between the different PRS pathways to indicate genes in common and potential biological overlap between pathways.

5.2 Materials and Methods

The PRS set-based method was applied to the AD data discussed in Section 2.1. Two independent datasets are required for the $PRS_{set-based}$ method, these are the test set containing individual genotype data and discovery data containing summary statistic in-

formation for each SNP. The set-based risk scores were produced for individuals in the GERAD data, this is the test set. The IGAP data excluding GERAD subjects was considered as the discovery set, using the summary statistics to improve power.

Since it is widely acknowledged that data should be pruned for LD prior to a PRS analysis [72]; correlation between SNPs was removed using intelligent pruning (--clump option in PLINK [50]). All SNPs which have not yet been clumped (index SNPs) and have p-values smaller than the set threshold were used to form clumps of all other SNPs within a 500kb distance which are in LD with the index SNP based on an r^2 threshold of 0.2 [50][51]. The p-value thresholds for both index SNPs and clumped SNPs were set to 1 to include all possible markers in the gene.

60,510 SNPs from a total of 419,048 total SNPs were retained from this clumping procedure in the genotyped data.

The HRC imputed GERAD data were additionally used for this analysis, again, these data were clumped in the same way as the genotyped data, retaining 193,369 SNPs of 6,119,694 total SNPs.

5.2.1 Gene-Based Analysis

The application of PRSs as a gene-based method in AD data was investigated.

The clumped SNPs in GERAD (both genotyped and imputed data separately) were assigned to genes using GENCODE (v19) gene models [78]. Only genes with known gene status and those marked as protein coding were used. SNPs which belong to multiple genes were assigned to all genes.

A total of 62,179 SNPs were assigned to 11,909 unique genes for the genotype data, where some SNPs are assigned to multiple genes. The HRC imputed data has 97,339 SNPs assigned to 12,218 unique genes.

A PRS was generated for each of these genes, and the gene-based PRS was tested for it's association with AD using a logistic regression model, adjusting for population covariates

such as age, sex, ethnicity and PCs to adjust for population stratification.

It was investigated whether the $PRS_{set-based}$ method was biased by the number of SNPs in the gene using the correlation between the $-log_{10}$(p-values) and the number of SNPs in the gene.

Finally, it was assessed whether genes identified are enriched in conserved regions. This was done both for genes which are evolutionary constrained and for genes in CNS which are less prone to variation [80]. Of the genes from the gene-based analysis, the number which were in conserved regions and the number of significant genes using a gene-wide significant p-value threshold of 2.5×10^{-6} [83] and a nominal threshold of 0.05 were determined. A chi-squared test was then used to determine whether an association between gene significance and whether the genes are in conserved regions exists, if cell counts are small then a Fisher's exact test is used in place of a chi-squared test.

5.2.2 Pathway Analysis

A PRS was computed for the set of SNPs which belong to each of the eight pathways determined as being associated with AD previously [23][24]. These pathways may aid in the understanding of different forms of AD and may regulate the early stages of the disease [91]. A self-contained test is determined by a logistic regression model with the pathway risk score, and a competitive analysis uses a likelihood ratio test to determine the added benefit of including the pathway risk score to a model including PRS across the whole genome.

The GERAD genotype data is used for the pathway analysis, and is again informed using summary statistics from IGAP data excluding the GERAD subjects.

Only SNPs with a p-value less than 0.5 in the IGAP study were used in the pathways. Each pathway PRS was generated, and the self-contained association of the pathway PRS with AD was found using a logistic regression model, adjusting for population covariates.

The association of each pathway with AD was assessed, and additionally, the pairwise

correlation between all pathways was investigated. A Pearson's correlation test between every pairwise combination of pathway PRSs was produced (using cor.test() in R software). It was expected that some correlation between pathways exists, since genes may belong to more than one biological pathway.

5.3 Results

5.3.1 Gene-Based Analysis

The PRS, MAGMA-PCA, Fisher and Simes gene-based methods were used in AD data in order to determine if any novel genes associated with AD could be found.

The results for the $PRS_{gene-based}$ method are seen in Table 5.1, the genes displayed are those which have previously been found to be associated with AD[19][20][21]. No novel genes were determined using the $PRS_{gene-based}$ method despite simulations showing the increased power compared to other gene-based methods when a real LD structure was simulated. This is likely due to the analysis being executed in clumped, genotyped data only, whereas the reported results were obtained using imputed data and combining two stages of IGAP [21]. Genes which have reached gene-wide significance replicate findings from [21] which uses Brown's method, the p-values from this analysis are also displayed in Table 5.1.

Table 5.1: PRS Gene-based Analysis in AD Genotype Data

			PRS			
Chr	Gene	No. of SNPs	β	SE	P-value	P-value from [21]
19	*PVRL2*	4	0.328	0.0303	2.9×10^{-27}	$< \times 10^{-300}$
19	*TOMM40*	1	0.313	0.0311	6.3×10^{-24}	NA
1	*CR1*	2	0.103	0.0293	4.4×10^{-4}	3.5×10^{-7}
2	*BIN1*	6	0.110	0.0300	2.7×10^{-4}	4.8×10^{-6}
6	*CD2AP*	4	0.059	0.0300	5.1×10^{-2}	8.0×10^{-6}
7	*EPHA1*	4	0.041	0.0301	1.7×10^{-1}	3.9×10^{-7}
8	*CLU*	2	0.133	0.0301	9.5×10^{-6}	1.2×10^{-108}
11	*PICALM*	4	0.048	0.0302	1.1×10^{-1}	1.2×10^{-8}
19	*ABCA7*	3	0.097	0.0297	1.2×10^{-3}	3.0×10^{-7}
8	*PTK2B*	5	0.061	0.0302	4.4×10^{-2}	1.3×10^{-4}
11	*SORL1*	8	0.105	0.0308	6.8×10^{-4}	6.7×10^{-5}
14	*SLC24A4*	15	0.034	0.0300	2.6×10^{-1}	1.0×10^{-3}
14	*RIN3*	22	0.061	0.0298	4.0×10^{-2}	5.6×10^{-1}
18	*DSG2*	3	-0.058	0.0299	5.3×10^{-2}	5.9×10^{-1}
2	*INPP5D*	15	0.071	0.0303	1.9×10^{-2}	NA
5	*MEF2C*	6	0.034	0.0300	2.5×10^{-1}	3.2×10^{-2}
7	*NME8*	7	-0.003	0.0300	9.3×10^{-1}	NA
7	*ZCWPW1*	2	0.066	0.0297	2.5×10^{-2}	8.3×10^{-1}
14	*FERMT2*	5	0.045	0.0294	1.3×10^{-1}	2.5×10^{-3}
20	*CASS4*	2	-0.014	0.0306	6.5×10^{-1}	4.2×10^{-3}

From Table 5.2, it is seen that results from Simes', Fisher's and MAGMA-PCA methods are clearly similar to those using the gene-based PRS method. Although the PRS method seems to have equivalent or smaller p-values for the majority of genes compared to MAGMA-PCA, Fisher and Simes. Previously published results may differ to these since they used imputed data and considered SNPs which may be outside of the gene. Only the PRS method provides the overall gene direction of effect, whereas the other methods simply provide a measure for the strength of association.

Table 5.2: Gene-based Analysis Comparison for PRS, MAGMA, Simes' and Fisher's Methods in AD Genotype Data

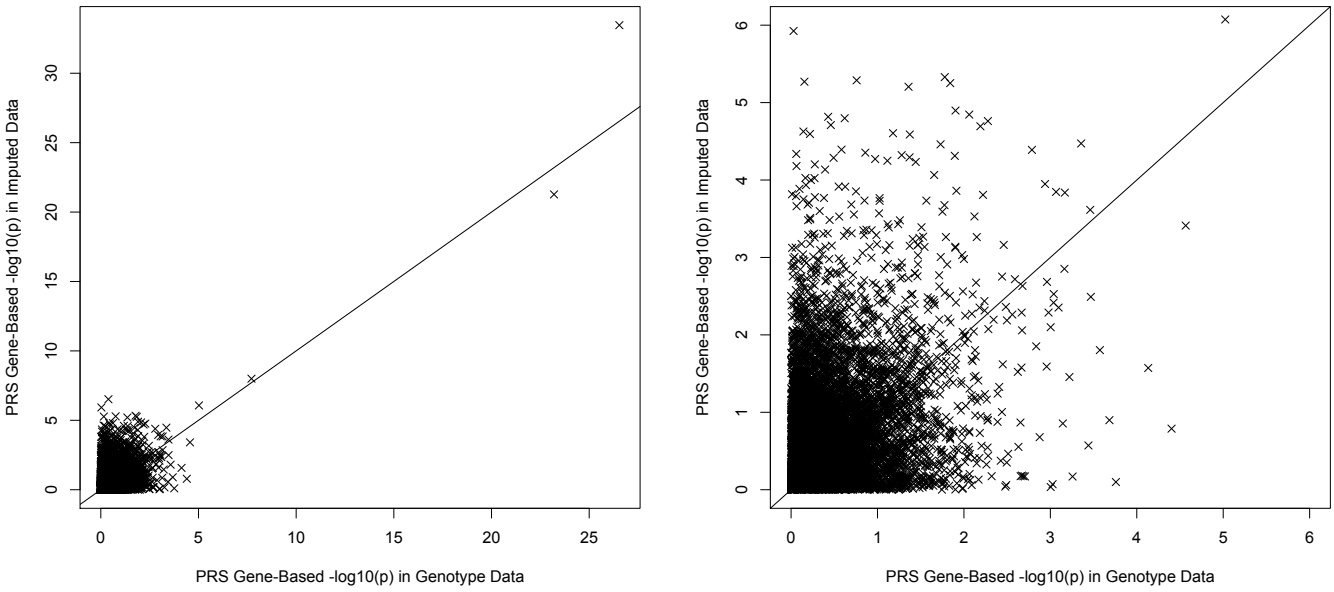
Chr	Gene	No. of SNPs	PRS P-value	Simes' P-value	Fisher's P-value	MAGMA-PCA P-value
19	*PVRL2*	4	2.9×10^{-27}	2.1×10^{-25}	$< 1.0 \times 10^{-16}$	4.8×10^{-25}
19	*TOMM40*	1	6.3×10^{-24}	6.4×10^{-24}	$< 1.0 \times 10^{-16}$	3.3×10^{-23}
1	*CR1*	2	4.4×10^{-4}	1.8×10^{-3}	4.4×10^{-4}	2.7×10^{-3}
2	*BIN1*	6	2.7×10^{-4}	3.9×10^{-3}	2.2×10^{-4}	1.9×10^{-3}
6	*CD2AP*	4	5.1×10^{-2}	2.8×10^{-1}	2.0×10^{-1}	4.1×10^{-1}
7	*EPHA1*	4	1.7×10^{-1}	1.4×10^{-1}	3.4×10^{-1}	4.7×10^{-4}
8	*CLU*	2	9.5×10^{-6}	1.0×10^{-5}	2.0×10^{-5}	10.0×10^{-6}
11	*PICALM*	4	1.1×10^{-1}	8.0×10^{-3}	2.1×10^{-3}	2.5×10^{-2}
19	*ABCA7*	3	1.2×10^{-3}	1.7×10^{-3}	2.0×10^{-3}	5.7×10^{-3}
8	*PTK2B*	5	4.4×10^{-2}	9.6×10^{-3}	4.1×10^{-2}	9.2×10^{-2}
11	*SORL1*	8	6.8×10^{-4}	4.4×10^{-2}	1.5×10^{-2}	8.7×10^{-2}
14	*SLC24A4*	15	2.6×10^{-1}	3.9×10^{-1}	1.6×10^{-1}	1.8×10^{-1}
14	*RIN3*	22	4.0×10^{-2}	2.2×10^{-1}	1.2×10^{-1}	5.4×10^{-1}
18	*DSG2*	3	5.3×10^{-2}	1.1×10^{-1}	3.0×10^{-1}	1.6×10^{-1}
2	*INPP5D*	15	1.9×10^{-2}	1.9×10^{-1}	1.7×10^{-3}	3.9×10^{-2}
5	*MEF2C*	6	2.5×10^{-1}	2.8×10^{-1}	2.7×10^{-1}	2.3×10^{-1}
7	*NME8*	7	9.3×10^{-1}	9.1×10^{-1}	9.8×10^{-1}	9.5×10^{-1}
7	*ZCWPW1*	2	2.5×10^{-2}	3.7×10^{-2}	6.0×10^{-3}	1.4×10^{-2}
14	*FERMT2*	5	1.3×10^{-1}	5.5×10^{-2}	3.8×10^{-2}	2.4×10^{-2}
20	*CASS4*	2	6.5×10^{-1}	5.8×10^{-1}	5.7×10^{-1}	5.5×10^{-1}

Table 5.3 displays the gene-wide significant ($p < 2.5 \times 10^{-6}$) genes from the gene-based PRS analysis in the imputed data. Three of the six genes are on chromosome 19 and are therefore likely influenced by *APOE* and *CLU* has also previously been identified as being associated with AD. Two additional genes on chromosomes 8 and 20, *CSMD1* and *MACROD2*, respectively, have been found to be associated with AD from this PRS gene-based analysis. In fact, *CSMD1* has been implicated in AD, familial Parkinson's disease [92] and cognitive performance [93]. The *MACROD2* gene has been implicated in neurological disorders [94].

Table 5.3: Gene-Wide Significant Genes from PRS Gene-based Analysis in AD Imputed Data

			PRS		
Chr	Gene	No. of SNPs	β	SE	P-value
8	*CSMD1*	579	0.191	0.0393	1.2×10^{-6}
8	*CLU*	1	1.540	0.3126	8.4×10^{-7}
19	*BCL3*	3	0.828	0.1445	1.0×10^{-8}
19	*PVRL2*	9	0.852	0.0699	3.4×10^{-34}
19	*TOMM40*	1	1.038	0.1076	5.4×10^{-22}
20	*MACROD2*	230	0.410	0.0801	3.0×10^{-7}

Figure 5.1 shows a plot of the $-log_{10}$(p-values) for the PRS gene-based analysis using genotype and imputed data. There are 10,243 genes in common between the gene-based analysis in the genotype and imputed data. It is clear there is a difference between the results from either type of data, particularly when considering the right graph in Figure 5.1, with 60% of genes having a smaller p-value from the imputed data compared to the genotype data. There is a positive correlation between the $-log_{10}$(p-values) for the imputed and genotype data (r=0.48, $p < 2.2 \times 10^{-16}$), and this correlation is weakened when the genes in Table 5.3 are removed (r=0.26, $p < 2.2 \times 10^{-16}$). This correlation demonstates that there is a difference between the gene-based results from the imputed and genotype data, since they are not perfectly correlated. This suggests that the gene-based analysis using the imputed data is more powerful, since it attains smaller p-values for a large number of genes, due to the inclusion of a larger number of SNPs.



(a) All common genes

(b) Genes with $-log_{10}(p) \leq 6$

Figure 5.1: Plot of Gene-Based $-log_{10}$(p-values) Using Genotype and Imputed GERAD Data

The imputed data includes a larger number of SNPs and therefore, it is expected that the number of SNPs in each gene is higher for the imputed data. A plot of this is shown in Figure 5.2.

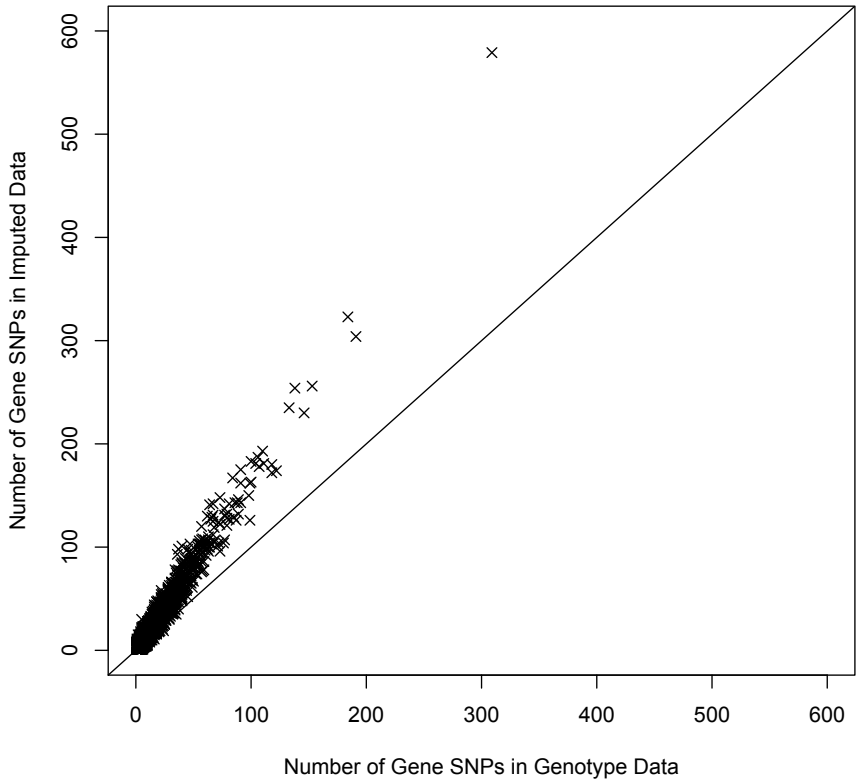


Figure 5.2: Plot of Number of Gene SNPs in Genotype and Imputed GERAD Data

5.3.1.1 Correlation Between P-values and the Number of SNPs in a Gene

Bias caused by gene size is an issue which is insufficiently tackled by most available gene-based methods [84][95]. This is because larger genes may harbour more significant SNPs by chance, and therefore the correlation between the number of SNPs per gene and the significance of a gene ($-log_{10}$(p-value)) in AD data was determined. SNP p-values may be inflated by population stratification and other confounders. This is okay when reporting the association of a single SNP with disease, but when combining a number of SNPs, the overall inflation can be substantial. The results can be seen in Table 5.4.

Table 5.4: Correlation Between $-log_{10}$(p-values) of Each Gene-Based Method and the Number of SNPs in the Gene

Gene-based Method	Correlation Coefficient	P-value
PRS (clumped)	-0.0054	0.555
PRS (unpruned)	-0.0105	0.204
MAGMA	0.0255	0.0021
Simes	0.0336	0.0003
Fisher	0.2144	$< 1.0 \times 10^{-16}$

Since there were a large number of genes with only 1 SNP (approximately 40%), the correlations were recalculated excluding these genes to ensure that results are not driven by these single SNP genes. Results are consistent when single SNP genes are excluded.

Clearly, the $PRS_{gene-based}$ method has no evidence of a correlation with the number of SNPs in a gene, suggesting that results are not inflated by gene size. However, all other methods show a correlation with the number of SNPs in the gene, the significant correlation coefficients are positive, indicating that a stronger association is observed when the set contains a larger number of SNPs, with there being very strong evidence of a correlation for Fisher's method.

The existence of a correlation between the gene size and the rate of false positives was considered using simulated data, where Nsim=1,000. The correlation between the type I error and the number of SNPs in the gene was determined for real data with permuted phenotypes to remove the effect sizes. The results are seen in Table 5.5. It is seen that there is no correlation between type I error and gene size for PRS and MAGMA-PCA methods. Fisher's method has a positive correlation, suggesting that increasing gene size increases the rate of false positive results. Simes' method shows an unusual result, it has evidence of a negative correlation, suggesting that an increase in gene size reduces the false positive rate. This is likely due to the fact that Simes' corrects for the number of SNPs in the gene.

Table 5.5: Correlation Between Power ($p \leq 0.05$) of Each Gene-Based Method and the Number of SNPs in the Gene

Gene-based Method	Correlation Coefficient	P-value
PRS	0.4697	0.4247
MAGMA-PCA (Disc+Test Genotype)	-0.1631	0.7932
MAGMA-PCA (Test Genotype)	-0.2199	0.7223
Simes	-0.9401	0.01742
Fisher	0.8734	0.05302

5.3.1.2 Conserved Regions

Loss of Function (LoF) genes from The Exome Aggregation Consortium (ExAC)

As discussed in Section 3.3.1.4, GWAS data was interrogated for genes that are evolutionary constrained, i.e, that are less likely to harbour variants of strong effect, probably due to functional importance. It was assessed whether there was enrichment for either LoF intolerant genes using the PRS gene-based results in both genotyped and imputed data.

Table 5.6 show the contingency tables at different p-value thresholds, 0.05 and 2.5×10^{-6} respectively, for LoF genes based on the PRS gene-based results in genotype data. For the 11,813 genes from the analysis, there is no evidence that the genes found here are constrained for both the nominal and gene-wide p-value threshold (p=0.0419 and p=0.4894 respectively). The Fisher's exact test is used for the gene-wide p-value threshold, since the cell counts are small.

Table 5.6: Number of LoF Genes from the Exome Aggregation Consortium, Genotype Data

	$p \leq 0.05$	$p > 0.05$
in LoF	112	2259
out LoF	550	8892

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	1	2370
out LoF	2	9440

(ii) $p \leq 2.5 \times 10^{-6}$

Since it is possible that some SNPs reside in multiple genes, it is not possible to assume that genes are independent, and the results in Table 5.6 are not adjusted for this correlation. Therefore, the analysis is repeated containing only non-overlapping genes, genes which were within 250kb of another gene were removed, with the most strongly associated gene retained. These results for the gene-based analysis in the genotype data are seen in Table

5.7. For the 555 non-overlapping genes, there is no evidence of an enrichment for genes in conserved regions at either the nominal or gene-wide p-value threshold (p=0.0937 and p=1 respectively).

Table 5.7: Number of LoF Genes from the Exome Aggregation Consortium, Genotype Data, No Overlapping Genes

	$p \leq 0.05$	$p > 0.05$
in LoF	43	64
out LoF	223	225

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	0	107
out LoF	1	447

(ii) $p \leq 2.5 \times 10^{-6}$

Similarly, the results for the 12,218 genes from the PRS gene-based analysis in imputed data are seen in Table 5.8. There is some evidence ($p=1.39 \times 10^{-6}$) at the nominal p-value threshold that the genes were in constrained regions, although there is no evidence (p=1) of an enrichment of genes in conserved regions at a gene-wide p-value threshold.

Table 5.8: Number of LoF Genes from the Exome Aggregation Consortium, Imputed Data

	$p \leq 0.05$	$p > 0.05$
in LoF	387	2066
out LoF	1182	8583

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	1	2452
out LoF	5	9760

(ii) $p \leq 2.5 \times 10^{-6}$

Table 5.9 shows the analysis in the imputed data was repeated for non-overlapping genes. Of the 519 genes remaining, there is no enrichment for genes in conserved regions at either the nominal or gene-wide threshold level (p=1 and p=0.5766 respectively).

Table 5.9: Number of LoF Genes from the Exome Aggregation Consortium, Imputed Data, No Overlapping Genes

	$p \leq 0.05$	$p > 0.05$
in LoF	91	41
out LoF	265	122

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	0	132
out LoF	4	383

(ii) $p \leq 2.5 \times 10^{-6}$

Conserved Noncoding Sequences (CNS)

The contingency tables showing whether the genes from the PRS gene-based analysis reside in CNS are seen in Table 5.10 for both nominal and gene-wide p-value thresholds. The cell counts are small for both tables and therefore a Fisher's exact test was used

to determine whether there was an enrichment of genes in conserved noncoding regions. There was no evidence of enrichment for either the nominal or gene-wide p-value threshold (p=1 for both).

Table 5.10: Number of Genes in Conserved Noncoding Sequences, Genotype Data

	$p \leq 0.05$	$p > 0.05$
in LoF	1	25
out LoF	661	11126

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	0	26
out LoF	3	11784

(ii) $p \leq 2.5 \times 10^{-6}$

Since genes may be correlated, the analysis was repeated considering only non-overlapping genes, results are shown in Table 5.11. Of the 555 non-overlapping genes, there is no evidence of an enrichment of genes in CNS at the nominal or gene-wide p-value threshold (p=0.2499 and p=1 respectively).

Table 5.11: Number of Genes in Conserved Noncoding Sequences, Genotype Data, No Overlapping Genes

	$p \leq 0.05$	$p > 0.05$
in LoF	0	3
out LoF	266	286

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	0	3
out LoF	1	551

(ii) $p \leq 2.5 \times 10^{-6}$

The same analysis was carried out for the 12,218 genes from the PRS gene-based analysis in the imputed data, the results are seen in Table 5.12. Again, a Fisher's exact test is used for both tables due to the small cell counts. No enrichment of genes in CNS is observed for the gene-wide p-value threshold (p=1), although there is some evidence (p=0.0015) of an enrichment at the nominal p-value threshold.

Table 5.12: Number of Genes in Conserved Noncoding Sequences, Imputed Data

	$p \leq 0.05$	$p > 0.05$
in LoF	9	16
out LoF	1560	10633

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	0	25
out LoF	6	12187

(ii) $p \leq 2.5 \times 10^{-6}$

For the 519 non-overlapping genes, there is also no enrichment of genes in CNS for either the nominal or gene-wide p-value threshold (p=1 for both), see Table 5.13.

Table 5.13: Number of Genes in Conserved Noncoding Sequences, Imputed Data, No Overlapping Genes

	$p \leq 0.05$	$p > 0.05$
in LoF	1	0
out LoF	355	163

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	0	1
out LoF	4	514

(ii) $p \leq 2.5 \times 10^{-6}$

There is no consistent enrichment of genes in conserved regions at the gene-wide significance level. It is expected that no enrichment would be seen, since AD is a post-reproductive disorder. Analyses considering non-overlapping genes are consistent, suggesting that the analysis is not biased by correlation between genes, or that the correlation is minimal.

5.3.2 Pathway Analysis

Eight pathways were found to be associated with AD using the ALIGATOR algorithm in the IGAP data [23][24]; these pathways are immune response, regulation of endocytosis, cholesterol transport, hematopoietic cell lineage, proteasome-ubiquitin activity, reactome hemostasis, clathrin and protein folding. The most strongly associated pathway with AD is the immune response pathway. The ALIGATOR algorithm [61] defines genes to be significant if they contain a single SNP with a p-value less than a set threshold, these gene sets are then compared to randomly generated gene sets, and as such, adjusts for gene size.

The use of PRS as a set-based method has been shown to be more powerful than other set-based methods, MAGMA-PCA, Fisher's and Simes (see Chapter 4 for details). The PRS approach can be applied to any set of SNPs, e.g. pathways. The PRS method provides an effect size for the strength and direction of association for each pathway, and produces a pathway risk score per person.

The PRS pathway scores were computed for the eight pathways previously found to be associated with AD [23][24]. The results are seen in Table 5.14 for both self-contained and competitive analyses. The self-contained test does not adjust for a baseline level of

association, whereas the competitive analysis does, by including the whole genome PRS in the model. The table also includes the p-values from the original ALIGATOR anaysis for comparison [61]. A pathway is considered as significant if the p-value is below 1.56×10^{-3}, using a Bonferroni correction $(0.05 \div (8 \times 4))$ [71]. It is seen that six of the eight pathways are associated with AD in the self-contained test; these are immune response, regulation of endocytosis, cholesterol transport, hematopoietic cell lineage, reactome hemostasis and clathrin. However, only the immune response and hematopoietic cell lineage pathways remain significant in the competitive test. ALIGATOR will be less influenced by noise, since it considers genes to be associated if they contain at least one associated SNP, whereas the PRS method includes all SNPs with a p-value less than 0.5. This explains why the ALIGATOR p-values are consistently lower than those from the pathways calculated using PRS.

Table 5.14: AD Associated Pathways Calculated Using PRS in GERAD data

Pathway	Beta	SE	P_{sc}	P_c	ALIGATOR p-value
1. Immune response	0.2695	0.0303	5.63×10^{-19}	4.75×10^{-8}	0.00266
2. Regulation of endocytosis	0.0993	0.0300	9.31×10^{-4}	0.282	0.0002
3. Cholesterol transport	0.1290	0.0305	2.37×10^{-5}	2.70×10^{-3}	0.00024
4. Hematopoietic cell lineage	0.1530	0.0298	2.84×10^{-7}	5.04×10^{-5}	0.00007
5. Proteasome-ubiquitin activity	0.0876	0.0303	3.89×10^{-3}	0.328	0.00929
6. Reactome hemostasis	0.1503	0.0300	5.47×10^{-7}	0.056	0.00785
7. Clathrin	0.1616	0.0305	1.19×10^{-7}	5.99×10^{-3}	0.00038
8. Protein folding	0.0949	0.0299	1.49×10^{-3}	0.088	0.00634

The analysis was also repeated removing the *APOE* region, since this has a large effect in AD, see Table 5.15. From the self-contained analysis, the same six pathways remain significant as in the analysis including *APOE*. However, with the exclusion of the *APOE* region, the immune response pathway is no longer significant in the competitive analysis but the hematopoietic cell lineage pathway remains significant. Note that the PRS pathway approach finds more consistently significant results in these eight pathways, compared

to the MAGMA approach, see Chapter 3.

Table 5.15: AD Associated Pathways Calculated Using PRS in GERAD data Excluding *APOE* Region

Pathway	Beta	SE	P_{sc}	P_c
1. Immune response	0.1644	0.0302	5.06×10^{-8}	0.048
2. Regulation of endocytosis	0.0993	0.0300	9.31×10^{-4}	0.282
3. Cholesterol transport	0.1011	0.0305	9.14×10^{-4}	0.036
4. Hematopoietic cell lineage	0.1448	0.0298	1.15×10^{-6}	1.23×10^{-4}
5. Proteasome-ubiquitin activity	0.0816	0.0303	0.007	0.410
6. Reactome hemostasis	0.1449	0.0300	1.39×10^{-6}	0.075
7. Clathrin	0.1604	0.0305	1.44×10^{-7}	0.006
8. Protein folding	0.0906	0.0299	0.002	0.110

The pairwise correlation between all pathways is considered, since this may suggest potential gene-overlap or the biological interaction between pathways. The pathway risk scores are adjusted for population stratification by regressing the scores against principal components, and testing the pairwise correlation between the residuals from this model. Table 5.16 shows the correlation between all of the pathway risk scores. The numbers in the table correspond to the numbers in Table 5.14. It is seen that the immune response pathway is highly correlated with all other pathways, possibly due to the impact of *APOE*. The cholesterol transport pathway is another which is highly correlated with the majority of other pathways. Although the correlations are highly significant, the actual correlation coefficients are fairly low, this is caused by the large sample size. Note, however, that all correlation coefficients are positive. The largest correlation coefficient is between the immune response pathway and the reactome hemostasis pathways.

Table 5.16: Correlations Between AD Associated Pathways, where Correlations were Calculated Using Individual PRS, where PRS is adjusted for population stratification *(Pathway Numbers Correspond to those in Table 5.14)*

Corr. Coeff (p-value)	1	2	3	4	5	6	7	8
1	1 (0)	0.1569 ($< 2.2 \times 10^{-16}$)	0.0895 ($< 2.2 \times 10^{-16}$)	0.1575 ($< 2.2 \times 10^{-16}$)	0.0979 ($< 2.2 \times 10^{-16}$)	0.3388 ($< 2.2 \times 10^{-16}$)	0.1412 ($< 2.2 \times 10^{-16}$)	0.0843 ($< 2.2 \times 10^{-16}$)
2		1 (0)	0.0323 (2.1×10^{-4})	0.0206 (0.0181)	0.0381 (1.2×10^{-5})	0.1072 ($< 2.2 \times 10^{-16}$)	0.1928 ($< 2.2 \times 10^{-16}$)	0.0369 (2.3×10^{-5})
3			1 (0)	0.0584 (2.0×10^{-11})	0.0602 (4.9×10^{-12})	0.0633 (3.7×10^{-13})	0.0879 ($< 2.2 \times 10^{-16}$)	0.1138 ($< 2.2 \times 10^{-16}$)
4				1 (0)	0.0210 (0.0161)	0.0818 ($< 2.2 \times 10^{-16}$)	0.0577 (3.4×10^{-11})	0.0366 (2.7×10^{-5})
5					1 (0)	0.0602 (4.9×10^{-12})	0.0489 (2.0×10^{-8})	0.0770 ($< 2.2 \times 10^{-16}$)
6						1 (0)	0.1668 ($< 2.2 \times 10^{-16}$)	0.0565 (8.9×10^{-11})
7							1 (0)	0.0544 (4.3×10^{-10})
8								1 (0)

5.4 Discussion

No novel genes have been determined using the PRS gene-based approach in the genotype data, although previous results have been replicated [21]. This is likely due to the use of LD pruned, genotyped data.

When compared to gene-based results using MAGMA-PCA, Simes or Fisher's methods, results from the PRS analysis have at least equivalent or more strongly associated p-values. These results are consistent with the simulation results in Chapter 4, where the $\text{PRS}_{set-based}$ method is used in simulated data with a real LD structure. This is expected, since the PRS approach utilises information from the additional discovery data to increase power.

The gene-based analysis in the GERAD imputed data determines two genes which have not previously been identified; *CSMD1* and *MACROD2*, both of which seem to be potentially interesting and biologically relevant candidates in AD. The gene-based analysis in the imputed data determined 6 gene-wide significant genes whereas only 2 gene-wide significant genes were found using the genotype data. This suggests the value in using imputed data which includes additional SNPs and thus can increase the power of the analysis.

Most set-based methods are biased by the number of SNPs in the set [84][95]. It is shown

that the $PRS_{set-based}$ method is not biased by the number of SNPs in the gene. This effect is demonstrated in MAGMA-PCA, Simes' and Fisher's methods. Although, this effect only holds for Fisher and Simes' methods when considering type I error, when the effect sizes of the SNPs are removed. This is expected for Fisher's approach, since it assumes independence between SNPs, and despite the fact that the data is pruned for LD, a threshold is still used and therefore some correlation will remain. Although Simes' also shows strong evidence of a correlation, the effect is in the opposing direction, i.e. association strength is increased when a smaller number of SNPs are in the set. This is likely because Simes corrects for the number of SNPs in the set, if this number is small, the correction will also be small, large sets are not an issue for Simes because it only considers whether one SNP in the set is associated with disease.

It was assessed whether genes from the gene-based analysis were in conserved regions, for genes which are evolutionary constrained and for genes which reside in CNS. No enrichment was observed for genes in regions of the genome which are evolutionary constrained, at a gene-wide significance threshold. As discussed previously, this is expected since AD is a post-reproductive disorder.

Six of the eight pathways previously found to be associated with AD [23][24] were additionally found to be associated using a self-contained test of the $PRS_{set-based}$ method. Only two pathways remained when a competitive test of association was used; these were the immune response and hematopoietic cell lineage pathways. Although the immune response pathway becomes non significant when the *APOE* region is excluded. These pathways have previously been tested with the early clinical manifestations of AD, in particular, the endocytosis pathway was found to be associated with AD, Mild Cognitive Impairment (MCI) and progression to dementia from cognitively normal [91]. The use of PRS pathways shows more consistent results with ALIGATOR compared to MAGMA pathways.

Many of the pathways show a correlation with one another, but the immune response pathway shows the strongest correlation with other pathways, this is potentially because it includes the *APOE* gene.

The $PRS_{set-based}$ method shows promise in this real data analysis, particularly when imputed data is used to improve power. The main disadvantage is that this approach does not account for LD and therefore requires LD pruning which substantially reduces the size of the data.

The $PRS_{set-based}$ method has the limitation in that it only provides a self-contained test of association, i.e. it does not take any baseline level of association into account. However, it would be possible to perform a competitive test of association which does takes account of the baseline level of association by either; including a whole genome PRS in the logistic regression model or using a simulation based approach matching SNPs by LD to find the empirical p-value.

This method can be extended to include any set of SNPs, as demonstrated here by considering both genes and pathways. It would also be possible to incorporate additional information into this score, such as gene expression in the brain, Deoxyribonucleic acid (DNA) methylation or other relevant biological features.

6 POLARIS: Polygenic LD-Adjusted Risk Score Set-Based Approach

6.1 Introduction

There would be some benefit in extending the standard PRS approach in order to adjust for LD between SNPs so that data would not require LD pruning prior to analysis; thus increasing the number of SNPs in the analysis and improving power (as discussed in Chapters 4 and 5).

In this chapter, a novel approach to a set-based framework which combines the advantages of MAGMAs PCA method and PRS is presented. The proposed POLARIS method [96] aims to improve upon the standard PRS method by correcting the inflated type I error observed both in standard PRS in the presence of LD [72], and also in set-based analyses as the number of SNPs in the set grows [84]. POLARIS uses spectral decomposition of the SNP correlation matrix to adjust the individuals' allele counts for LD structure. POLARIS is presented as a self-contained set-based approach in that it compares the test statistic for the set with the null hypothesis, rather than a competitive approach which accounts for the baseline level of association across the genome. However, it can be turned into a competitive method either by including a general whole genome POLARIS in the analysis, or comparing the set-based POLARIS to those generated from random sets of genes (matched for the number of SNP-sets/set size/number of SNPs). A self-contained test of association is beneficial to find the individual risk scores of each set which can be used in further analyses, however, the strength of the association may be inflated by

the baseline level of association across the genome. Therefore, a competitive test has the advantage that it gives a more accurate measure of overall set association since it accounts for the baseline association.

POLARIS informs the analysis with previously reported effect sizes of the SNP's association with disease. An LD-adjusted PRS is calculated per person per set, and the overall set effect is computed using regression. Since the score is used as a predictor in a regression analysis, it is possible to include further population covariates or any other possible confounders. POLARIS uses all available information, since all PCs are incorporated into a score, thus avoiding overfitting which may result from only including the top PCs. As in standard PRS analysis, only one independent variable (apart from extra covariates) is present in the regression model, rather than the number of predictors being equal to the number of markers, or the number of chosen PCs. Like the standard PRS approach, an advantage of this method is that it performs a self-contained test of association in the test dataset, leveraging the discovery set to increase the power of this test. A significant test statistic implies significant association specifically in the test sample, unlike a significant meta-analysis result, where the association evidence could result from other samples. This might be important if the test sample is of specific interest, e.g. a different ethnicity, or a different, but related, phenotype.

6.1.1 Objectives

This Chapter aims to:

- Develop a methodology which adjusts standard PRS for LD between SNPs called POLARIS.
- Assess the type I error and power of POLARIS in simulated data with constructed and real LD structure.
- Compare the power of POLARIS to that of the regression based approach in MAGMA which computes principal components for all SNP genotypes and uses an F-test to find the strength of association between the phenotype and the SNPs (MAGMA-

PCA). Also compare the power of POLARIS to that of the MAGMA mean-χ^2 approach which is used when only summary statistic data is available (MAGMA-SUMMARY) [55]. POLARIS is not compared to Fisher's and Simes methods since they do not adjust for LD.

6.2 Materials and Methods

6.2.1 POLARIS Rationale and Derivation

For M SNPs in a set, the standard PRS combines single-SNP genotypes g_i $(i = 1, \ldots, M)$ into a single regression predictor using single-SNP effect sizes ($\log(\text{OR}_i) = \beta_i$) taken from a previous study as coefficients (see also Chapter 4),

$$PRS = \sum_{i=1}^{M} \beta_i g_i = \beta^T g. \tag{6.1}$$

This method implements a 2-stage approach, where independent discovery and test sets are available. The effect sizes β are determined from the discovery set and a vector of the number of risk alleles g is obtained from the test set. The underlying assumption is that individual genotypes are available for the test set, but only summary data (effect sizes β) for the discovery set are available.

The standard PRS method does not adjust for LD between markers and thus requires LD pruning [72]. If markers are in LD, the simple weighted sum (Equation 6.1) may give them undue weight; indeed, if they are in positive LD, they are likely to have a similar single-SNP effect size and act together, thus giving a larger contribution to the PRS than a single or uncorrelated marker.

This imbalance due to LD is corrected by replacing the vector g of genotypes with a vector $\tilde{g}$ of adjusted dosages. Consider the spectral decomposition of the $M \times M$ marker-marker correlation matrix C,

$$C = \sum_{k=1}^{M} \lambda_k \, x_k \, x_k^T \tag{6.2}$$

with eigenvalues λ_k satisfying $\sum_{k=1}^{M} \lambda_k = \operatorname{tr} C = M$ and orthonormal (column) eigenvectors x_k, where $\operatorname{tr} C$ is the trace (sum of the elements on the main diagonal) of matrix C. The correlation matrix is the covariance matrix of the joint distribution of individual genotypes after standardisation of each SNP. Its eigenvectors indicate the directions of the principal axes of this standardised distribution, and the corresponding eigenvalues give the variances of the distribution in the corresponding directions. In the absence of LD, these variances will be equal to 1, and the distribution will be isotropic. However, if there is LD, then these variances will in general be different, and the standardised distribution will be more elongated in some principal directions and flattened in others.

This anisotropy can be removed by scaling the standardised joint distribution in the direction of each principal axis with the inverse square root of the eigenvalue in this direction. However, adjusting the standardised distribution in this way will not only remove LD, but also equalise the single marker variances, thus discarding information such as the MAFs. As our aim is to adjust for LD only, but not for single-SNP variances, the same scaling transformation is applied to the original, unstandardised joint distribution instead.

More specifically, due to the orthonormality of the eigenvectors, the PRS can be expressed in a spectral decomposition

$$PRS = \beta^T g = \sum_{k=1}^{M} \beta^T x_k \, x_k^T g. \tag{6.3}$$

The component $x_k \, x_k^T g$, which is the part of g along the kth principal axis, has correlation matrix eigenvalue λ_k and therefore contributes a disproportionate amount of variance to PRS unless $\lambda_k \approx 1$. For an uncorrelated marker, one spectral component will be concentrated on this marker, and the corresponding eigenvalue $\lambda_k \approx 1$.

For our adjustment, the coordinate of g is rescaled in the direction of the kth principal axis, $x_k^T g$, with the inverse square root of the correlation eigenvalue, giving an adjusted coordinate $\frac{1}{\sqrt{\lambda_k}} x_k^T g$, and hence the rescaled spectral component $\frac{1}{\sqrt{\lambda_k}} x_k \, x_k^T g$.

Applying this adjustment to each principal axis will result in an isotropic distribution in

which the correlation has mostly been removed, for the adjusted dosage vectors

$$\tilde{g} = \sum_{k=1}^{M} \frac{1}{\sqrt{\lambda_k}} (x_k^T g) = C^{-\frac{1}{2}} g. \tag{6.4}$$

Note this adjustment of multivariate data by correlation is analogous to the calculation of the Mahalanobis distance for mean zero data x, $x^T S^{-1} x = \|S^{-\frac{1}{2}} x\|^2$, where S is the covariance matrix [97][98], except that here the correlation matrix is used instead of the covariance matrix in order to avoid adjusting for single-marker variance.

Using the adjusted dosages $\tilde{g}$ instead of the the original genotype vectors g, obtains an LD-adjusted Polygenic Risk Score (PRS)

$$\sum_{i=1}^{M} \beta_i \tilde{g}_i = \beta^T \tilde{g} = \beta^T C^{-\frac{1}{2}} g = \sum_{i=1}^{M} \beta_i \left(\sum_{k=1}^{M} \frac{1}{\sqrt{\lambda_k}} x_k(i) \sum_{j=1}^{M} x_k(j)\, g_j \right). \tag{6.5}$$

In the sum over the spectral components, indexed by k, the terms with $\lambda_k = 0$, corresponding to principal directions with no variance, are to be omitted, resulting effectively in a pseudoinverse of the square root of C. In cases of extreme LD, where $\lambda_k \approx 0$, this formula will apply a large correction factor to the corresponding component, thus possibly amplifying small deviations due e.g. to genotyping error. In order to avoid this instability, a ridge parameter λ_0, is introduced, where $\lambda_0 = \sqrt{\frac{1}{N}}$, where N is the number of individuals in the test data, and the adjustment is modified to mitigate the effect of small λ_k. This gives rise to the POLARIS risk score,

$$POLARIS = \beta^T \sqrt{1+\lambda_0}\, (C + \lambda_0 I)^{-\frac{1}{2}}\, g = \sum_{i=1}^{M} \beta_i \left(\sum_{k=1}^{M} \sqrt{\frac{1+\lambda_0}{\lambda_k + \lambda_0}}\, x_k(i) \sum_{j=1}^{M} x_k(j) g_j \right) = \beta^T \tilde{g}, \tag{6.6}$$

where now $\tilde{g} = \sqrt{1+\lambda_0}\, (C + \lambda_0 I)^{-\frac{1}{2}}\, g = \sum_{k=1}^{M} \sqrt{\frac{1+\lambda_0}{\lambda_k+\lambda_0}} x_k\, x_k^T g$, and I is the $M \times M$ unit matrix. Note that if all markers are uncorrelated, then $\lambda_k \approx 1$ for all k, which makes $\tilde{g} \approx g$, and consequently $POLARIS \approx PRS$.

The adjustment $\tilde{g} = C^{-\frac{1}{2}} g$ (or the extension with a ridge parameter) is applied directly to the vector of genotypes. More precisely, an adjustment of the variance only will be

achieved by removing the sample mean vector $\hat{g}$ before the adjustment, giving

$$\tilde{g} = \hat{g} + C^{-\frac{1}{2}}(g - \hat{g}) = C^{-\frac{1}{2}}g + (I - C^{-\frac{1}{2}})\hat{g}; \tag{6.7}$$

however, this only amounts to shifting the POLARIS score by a constant $\beta^T(I - C^{-\frac{1}{2}})\hat{g}$, which is irrelevant in the subsequent regression analysis. Similarly, this would be irrelevant if the POLARIS score was normalised.

6.2.2 POLARIS Set-Based Analysis Comparison Applied to Simulated Data

The POLARIS methodology is compared to two different MAGMA approaches [55] in order to assess whether POLARIS has at least equivalent power compared with these approaches. The first is a regression based approach which computes the principal components for SNP genotypes, and determines how strongly associated these SNPs are with the phenotype of interest using an F-test (MAGMA-PCA). The second MAGMA approach is designed for when only summary statistic data are available, it uses a mean-χ^2 method and permutations to account for LD (MAGMA-SUMMARY), similar to Brown's method [52]. Of the set-based methods introduced in Chapter 4, only MAGMA-PCA and MAGMA-SUMMARY are considered here as these are the only approaches which also account for LD between SNPs.

To understand detailed differences and similarities between MAGMA approaches and POLARIS, all methods were tested on simulated data, both with a simple extreme constructed LD pattern and a real-data LD pattern between SNPs. Type I and type II errors were tested by simulating null effects and introducing some association to the SNPs, respectively.

To generate summary statistic data and genotype data, a simulated dataset was randomly split into discovery and test sets. The summary statistics for each SNP in the discovery set were computed. The following different scenarios were simulated.

- **Simple LD Block:** 10 SNPs in an LD Block with (i) $r^2 = 0.2$ and (ii) $r^2 = 0.8$

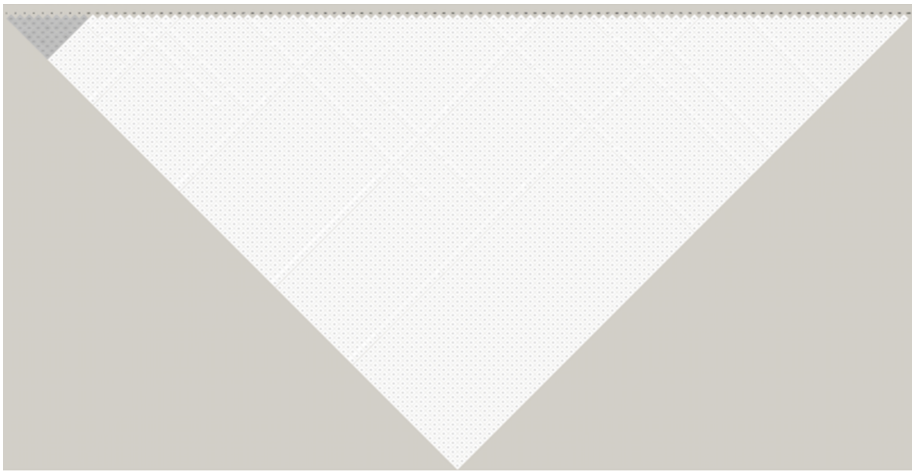
between consecutive SNPs. The 'causal' SNP is associated with disease with OR=1.1 and the remaining 9 SNPs have an OR closer to the null value of 1. An additional 90 independent unassociated SNPs are also present in the set, see Figure 6.1 for LD structure.

- **Complex LD:** 4 LD Blocks of 10 SNPs each, and 60 independent unassociated SNPs. Block 1 has pairwise $r^2 = 0.2$, Block 2 has pairwise $r^2 = 0.4$, Block 3 has pairwise $r^2 = 0.6$, and Block 4 has pairwise $r^2 = 0.8$, all 40 SNPs in LD with OR $\sim$ N(1.02, 0.36) (OR from a Normal Distribution with mean 1.02 and variance 0.36), see Figure 6.2 for LD structure. The mean and variance for the sampled effect sizes are calculated from all SNPs in the IGAP data [20].

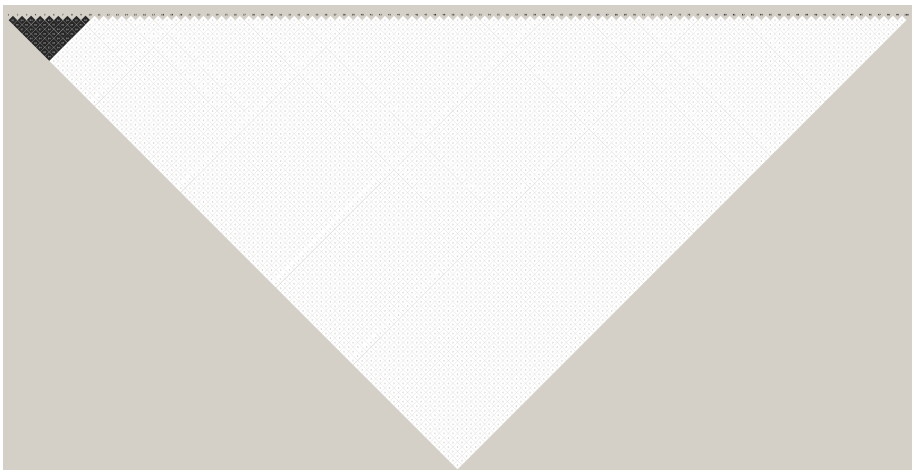
- **Discovery and Test with Different LD Structure:** 10 SNPs in LD with OR$\sim$ N(1.02, 0.36) and 90 independent, unassociated SNPs where test set LD is moderate ($r^2 = 0.6$) and discovery set LD is high ($r^2 = 0.8$), see Figure 6.3 for the LD structure.

- **Effect Sizes of Varying Direction:** It is possible that for certain MAFs, SNPs in LD have effects in opposite directions [88]. 10 SNPs with varying LD with ORs with randomly varying direction and 90 independent unassociated SNPs. The LD structure can be seen in Figure 6.4.

- **Real Data Simulations:** 115 SNPs from real AD GERAD data [19], see Section 2.1 for a detailed description of the data and Figure 6.5 for LD structure. For a SNP in a block of strong LD, a number of controls who were homozygous for the risk allele were set to cases, and an equal number of cases homozygous for the protective allele were set to controls, thus producing an association with disease.



(a) $r^2 = 0.2$



(b) $r^2 = 0.8$

Figure 6.1: LD Plot for 100 SNPs in Simple LD Simulations

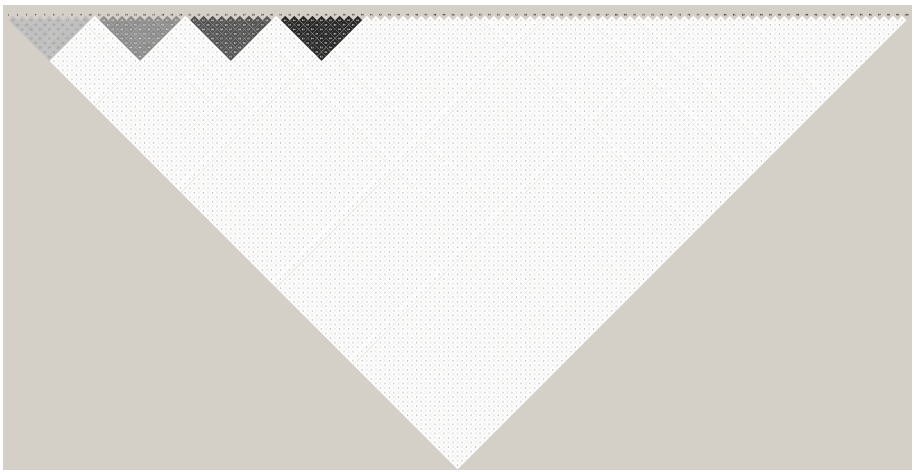
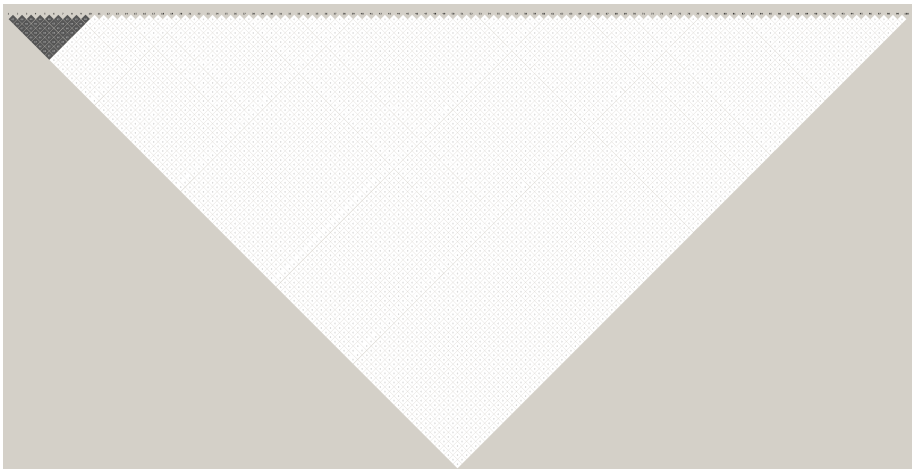
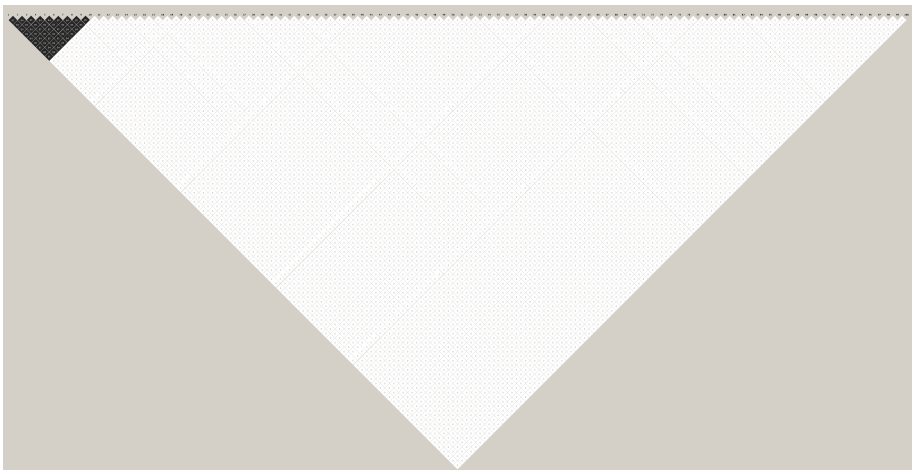


Figure 6.2: LD Plot for 100 SNPs in Complex LD Simulations



(a) Test Set $r^2 = 0.6$



(b) Discovery Set $r^2 = 0.8$

Figure 6.3: LD Plot for 100 SNPs in Discovery and Test with Different LD Structure Simulations

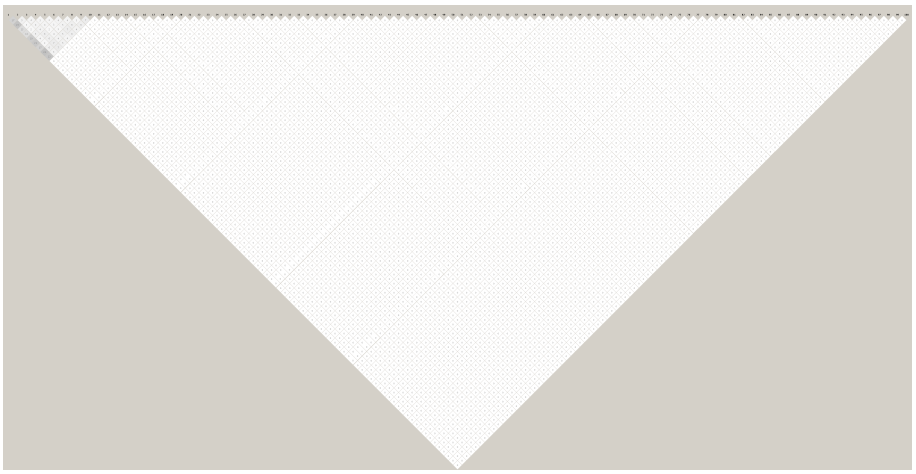


Figure 6.4: LD Plot for 100 SNPs with Varying Effect Sizes Simulation

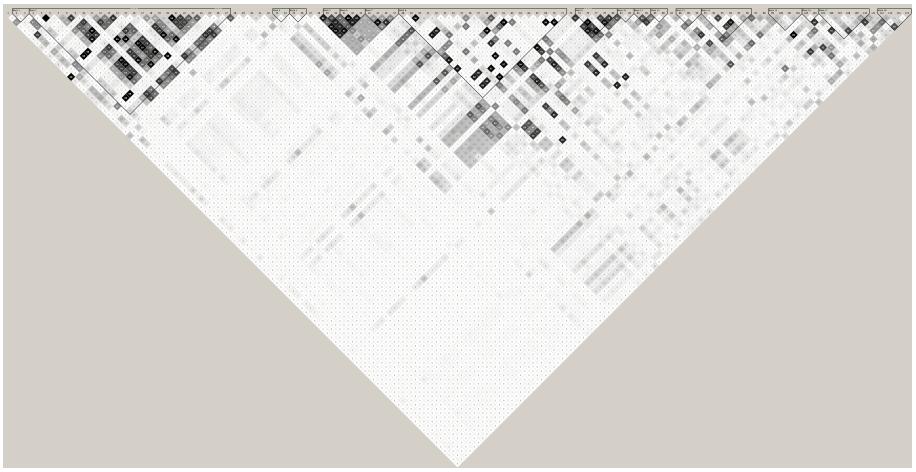


Figure 6.5: LD Plot for Real LD Structure with 115 SNPs

The real data simulations in this chapter use a smaller number of SNPs compared to the same simulations in Chapter 4 (115 SNPs compared to 129 SNPs respectively). The reason for this is that only SNPs with a p-value greater than 0.1 were included; this was in an attempt to remove some of the effects in the data so it was possible to better distinguish between the power of the different approaches.

For these scenarios, the sample size of the discovery dataset was varied in order to determine the influence of the discovery set sample size on the POLARIS method. Simulations were run creating data with N=20,000 and N=60,000 individuals, these were split equally to result in a test and discovery set each with N=10,000 and N=30,000 subjects, respectively. Additionally, the larger set with 60,000 individuals was split such that the test set had N=10,000 and the discovery set had N=50,000 individuals. The real data simulations have 13,164 subjects each for the combined discovery and test sets; these data are divided 50/50 and 25/75. In both the discovery and test datasets, 30% of the sample size were cases.

A total of 1,000 simulations were performed for each scenario. The power to detect the association between the set and disease is calculated as the proportion of p-values from the 1,000 simulations which were below a given p-value threshold; the p-value thresholds used were p=0.05, 0.01 and 0.001. 10,000 simulations were used for the real data simulations, thus enabling a more stringent threshold of 0.0001 to be considered. The power of the POLARIS method, applied to the test dataset and informed by the discovery dataset, was

compared to the power of MAGMA-PCA applied to the test dataset only, MAGMA-PCA to the total unsplit data, MAGMA-SUMMARY using discovery set summary statistics and test data to estimate LD and MAGMA-SUMMARY in the combined test and discovery sets.

It was also investigated whether the adjustment using the square inverse of the correlation matrix was better in terms of power than simply using the inverse of the correlation matrix. This was the initial approach investigated until the Mahalanobis distance theory was considered (see Section 6.2.1), so this comparison is to confirm the theoretical equation with simulations. A type I error and power comparison between the square inverse and inverse of the correlation matrix was considered for the simple, complex and real LD structure simulations.

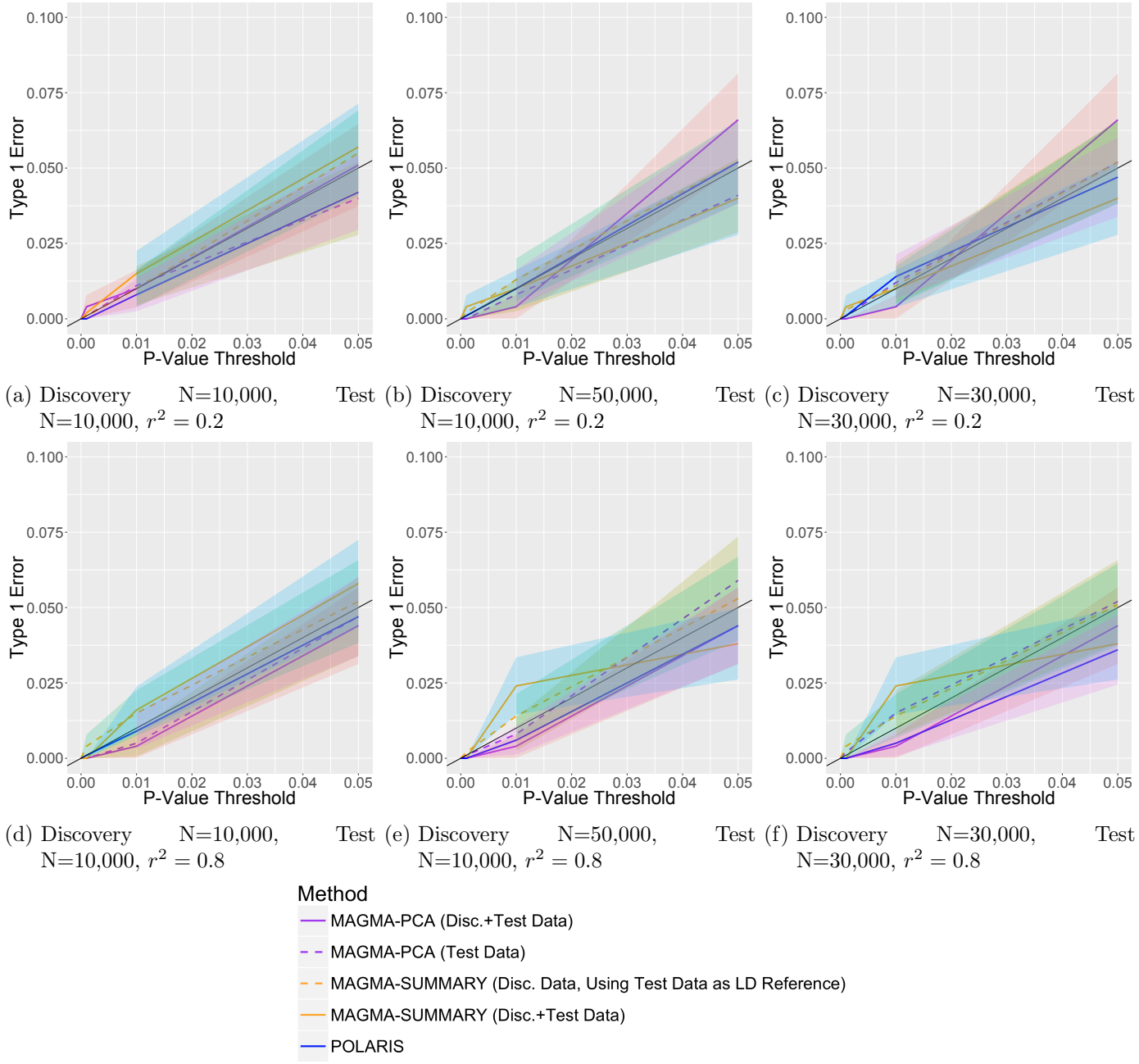
6.3 Results

6.3.1 Type I error

All scenarios outlined above in Section 6.2.2 were tested for type I error rates, where none of the SNPs have an association to disease (i.e. OR=1) in either the discovery or test sets. Type I error is deemed acceptable if the nominal value is included in the 95% CI for estimated type I error rate. The expected type I error is displayed on the type I error plots (black solid line). The POLARIS results are shown by a solid blue line on the plots. MAGMA-PCA is displayed as a purple line; the solid purple line is in the test and discovery data and the dashed purple line is in the test data only. MAGMA-SUMMARY is shown as an orange line; the solid orange line is in the combined test and discovery data and the dashed orange line is in the discovery set only, using the test set to estimate LD. The shaded area surrounding each line displays the 95% CI for the type I error for each method.

6.3.1.1 Simple LD Block

Figure 6.6 shows the type I error for the Simple LD simulation example. 10 SNPs out of 100 are in LD with either $r^2 = 0.2$ or $r^2 = 0.8$. The LD structure for these simulations can be seen in Figure 3.7a. The 95% CIs are also displayed on the figures. The type I error rate is reasonable in the majority of cases; the nominal value is included in the 95% CI.



(a) Discovery N=10,000, Test N=10,000, $r^2 = 0.2$

(b) Discovery N=50,000, Test N=10,000, $r^2 = 0.2$

(c) Discovery N=30,000, Test N=30,000, $r^2 = 0.2$

(d) Discovery N=10,000, Test N=10,000, $r^2 = 0.8$

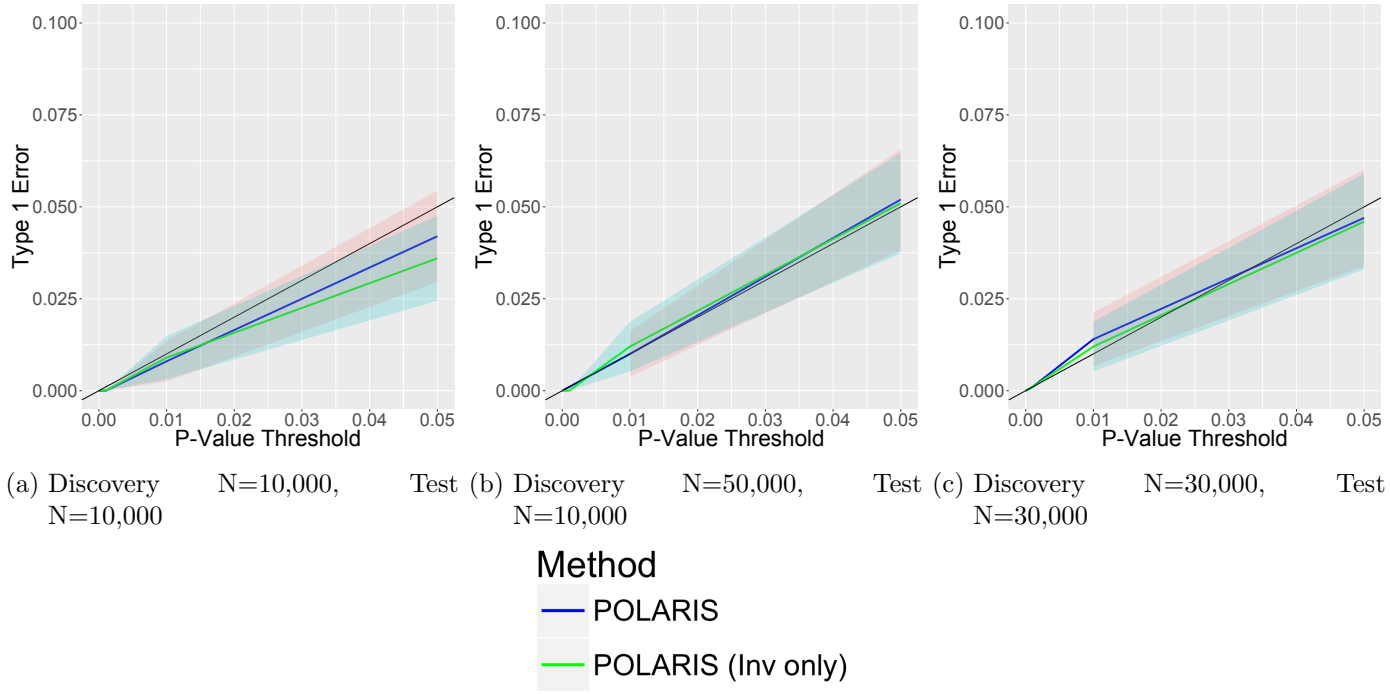
(e) Discovery N=50,000, Test N=10,000, $r^2 = 0.8$

(f) Discovery N=30,000, Test N=30,000, $r^2 = 0.8$

Figure 6.6: Type I Error Comparison of Set-Based Methods; Simulation of 10 SNPs in LD and 90 independent SNPs. Figures 6.6a, 6.6b and 6.6c show Simple LD Structure Simulations where $r^2 = 0.2$, and Figures 6.6d, 6.6e and 6.6f show Simple LD Structure Simulations where $r^2 = 0.8$. Figures 6.6a and 6.6d have a discovery and test sample size of 10,000, Figures 6.6b and 6.6e have a discovery set N=50,000 and test set N=10,000 and Figures 6.6c and 6.6f have discovery and test sets with N=30,000. *Note: y-axis scale is not between 0 and 1.*

The comparison between POLARIS which uses the square inverse of the correlation matrix to adjust for LD (blue solid line), and an adjustment using the inverse of the correlation matrix only (green solid line) is shown in Figure 6.7. The expected type I error is shown by

the black solid line. It is seen that the type I error for both adjustments are comparable.

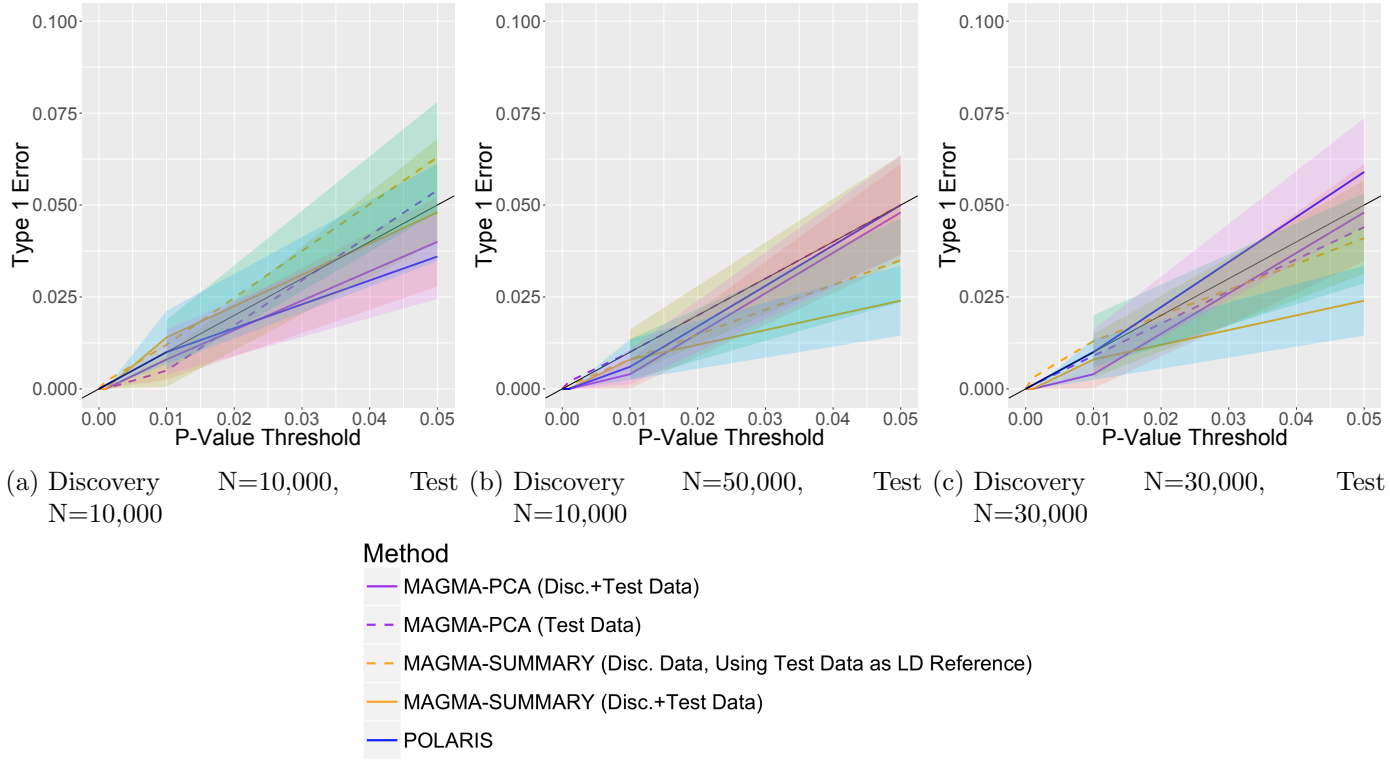


(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 6.7: Type I Error Comparison of POLARIS and an Adjustment using the Inverse of the Correlation Matrix; Simulation of 10 SNPs in LD and 90 independent SNPs. *Note: y-axis scale is not between 0 and 1.*

6.3.1.2 Complex LD Structure

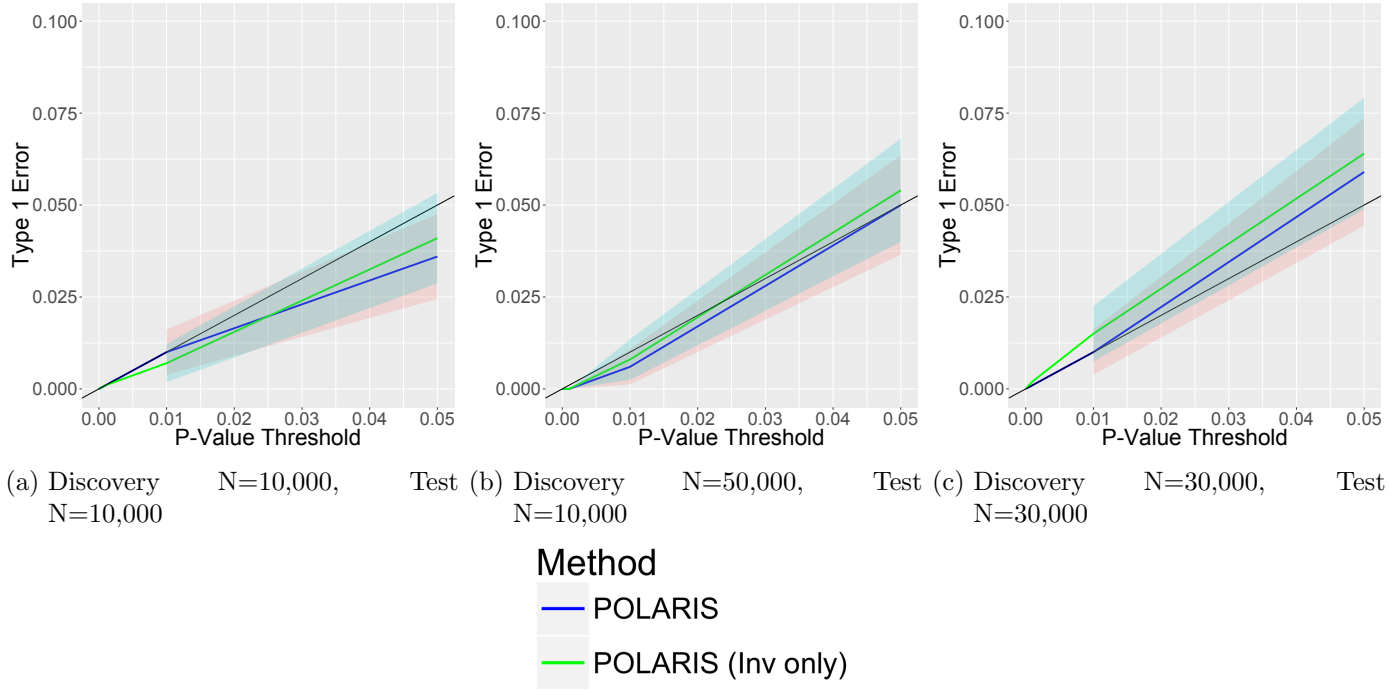
The type I error for the simulation with complex LD structure is shown in Figure 6.8. The first 10 SNPs have pairwise LD $r^2 = 0.2$, the second 10 SNPs have pairwise LD $r^2 = 0.4$, the third 10 SNPs have pairwise LD $r^2 = 0.6$ and the fourth 10 SNPs have pairwise LD $r^2 = 0.8$. The remaining 60 SNPs are independent. The nominal value is within the 95% CI and therefore type I error is reasonable.



(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 6.8: Type I Error Comparison of Set-Based Methods; Simulation of 10 SNPs in LD with $r^2 = 0.2$, 10 SNPs in LD with $r^2 = 0.4$, 10 SNPs in LD with $r^2 = 0.6$, 10 SNPs in LD with $r^2 = 0.8$ and 60 independent, unassociated SNPs. *Note: y-axis scale is not between 0 and 1.*

The comparison between the standard POLARIS LD adjustment and the adjustment using the inverse of the correlation matrix for this simulation is seen in Figure 6.9. Again, the type I error for both adjustments is similar.

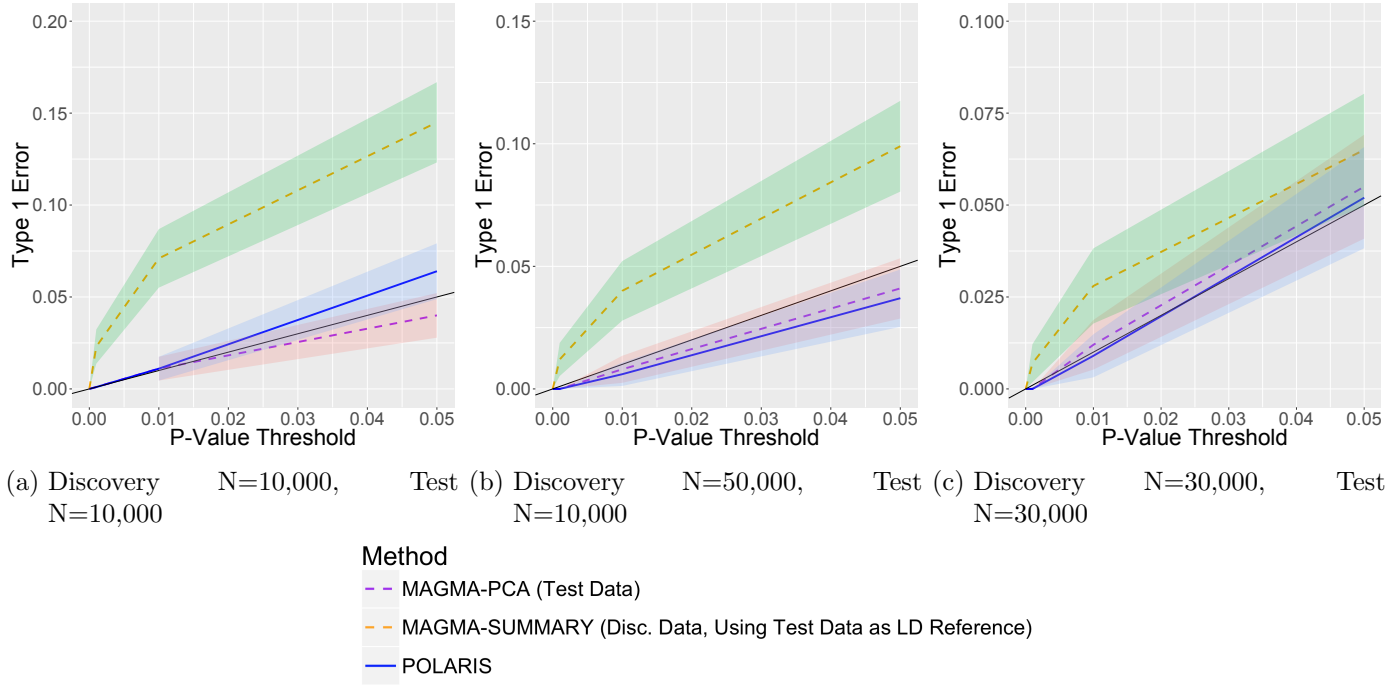


(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 6.9: Type I Error Comparison of POLARIS and an Adjustment using the Inverse of the Correlation Matrix; Simulation of 10 SNPs in LD with $r^2 = 0.2$, 10 SNPs in LD with $r^2 = 0.4$, 10 SNPs in LD with $r^2 = 0.6$, 10 SNPs in LD with $r^2 = 0.8$ and 60 independent, unassociated SNPs. *Note: y-axis scale is not between 0 and 1.*

6.3.1.3 Different LD Structure of Discovery and Test Datasets

Figure 6.10 shows the type I error for the simulations which have a different strength of LD between the test and discovery sets, both sets have 10 SNPs in LD and the remaining SNPs are independent. The pairwise LD in the test set has pairwise $r^2 = 0.6$ and the discovery set has pairwise LD $r^2 = 0.8$. Type I error is inflated for MAGMA-SUMMARY in the discovery data only (dashed orange line), this is due to the fact that LD is estimated from a test set which has a different LD structure to the discovery set, and so the discovery set is under-adjusted for LD. MAGMA-PCA and MAGMA-SUMMARY in the combined data are not presented here, since the combined data will have an average LD and would therefore not be comparable.



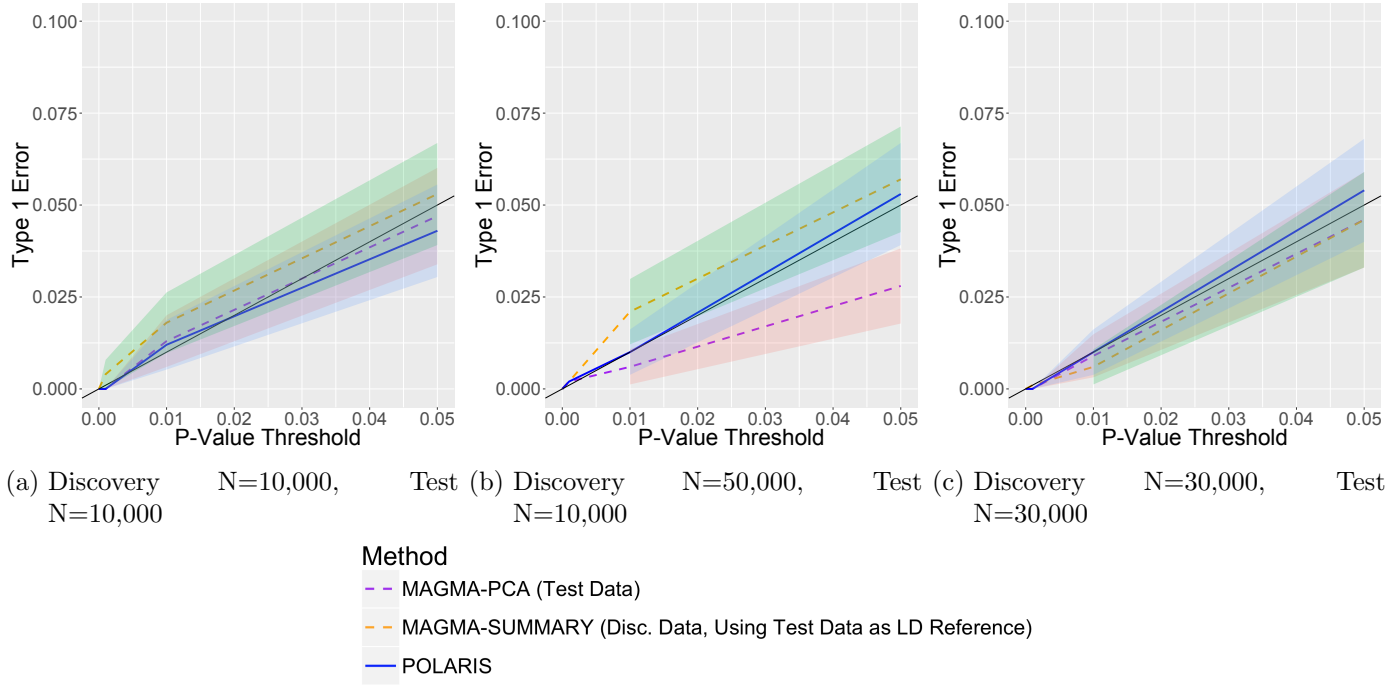
(a) Discovery N=10,000, Test N=10,000

(b) Discovery N=50,000, Test N=10,000

(c) Discovery N=30,000, Test N=30,000

Figure 6.10: Type I Error Comparison of Set-Based Methods; Simulation of 10 SNPs in LD with OR$\sim$ N(1, 0.2^2) and 90 independent, unassociated SNPs where Test LD is moderate ($r^2 = 0.6$) and Discovery LD is high ($r^2 = 0.8$). *Note: y-axis scale is not between 0 and 1.*

When considering the opposite case; where the test set has high LD ($r^2 = 0.8$) and the discovery set has moderate LD ($r^2 = 0.6$), see Figure 6.11. The type I error is no longer inflated for the MAGMA-SUMMARY method, since the LD in the test set which is used to adjust, is larger than the LD in the discovery set. Therefore, the adjustment will be more severe in this case.

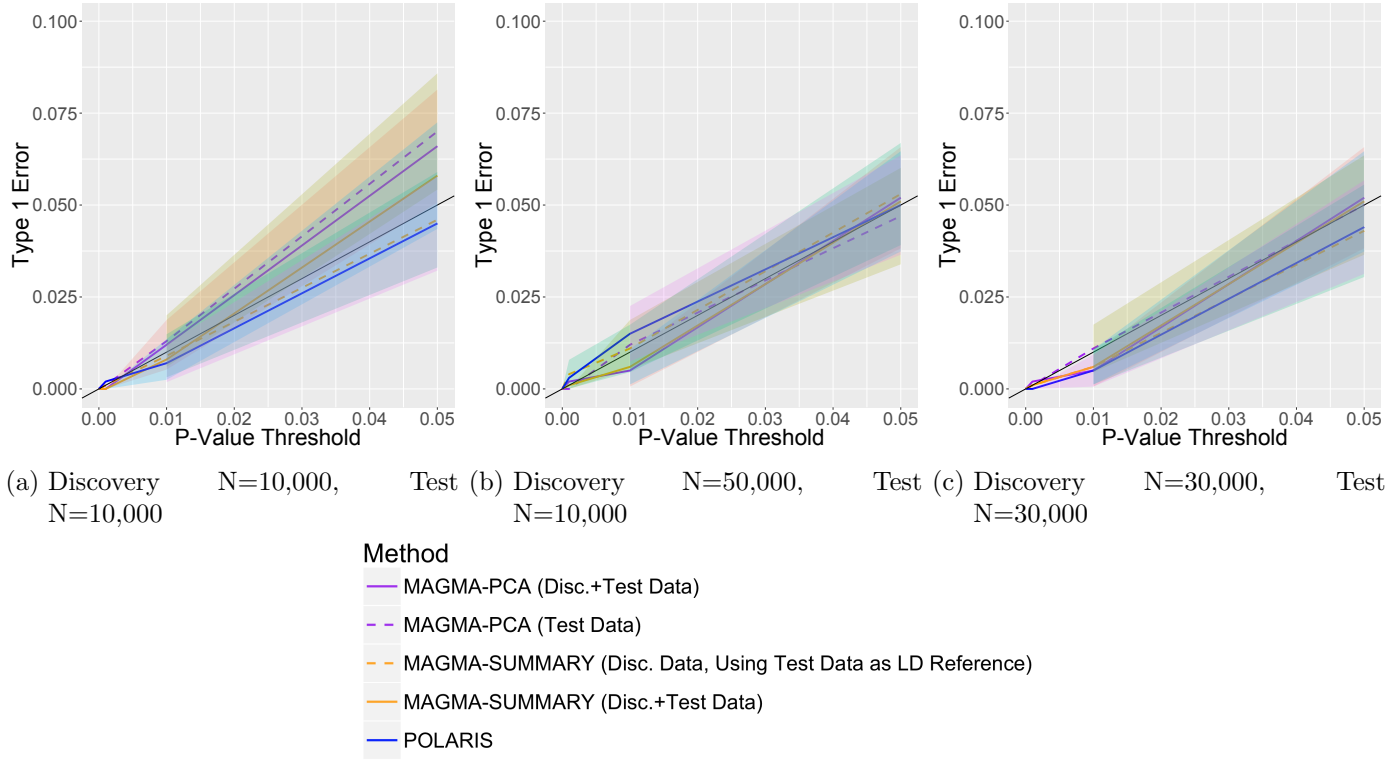


(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 6.11: Type I Error Comparison of Set-Based Methods; Simulation of 10 SNPs in LD with OR$\sim$ N(1, 0.2^2) and 90 independent, unassociated SNPs where Test LD is high ($r^2 = 0.8$) and Discovery LD is moderate ($r^2 = 0.6$). *Note: y-axis scale is not between 0 and 1.*

6.3.1.4 Effect Sizes with Varying Direction

The type I error for the simulation where the association of the SNPs in the LD block are not all in the same direction is shown in Figure 6.12. Although, for the type I error case, there are no effect sizes so the direction of effect can not vary. Therefore, this case is similar to the simple LD case with pairwise $r^2 = 0.2$. As before, type I error is within the 95% CIs for all methods.

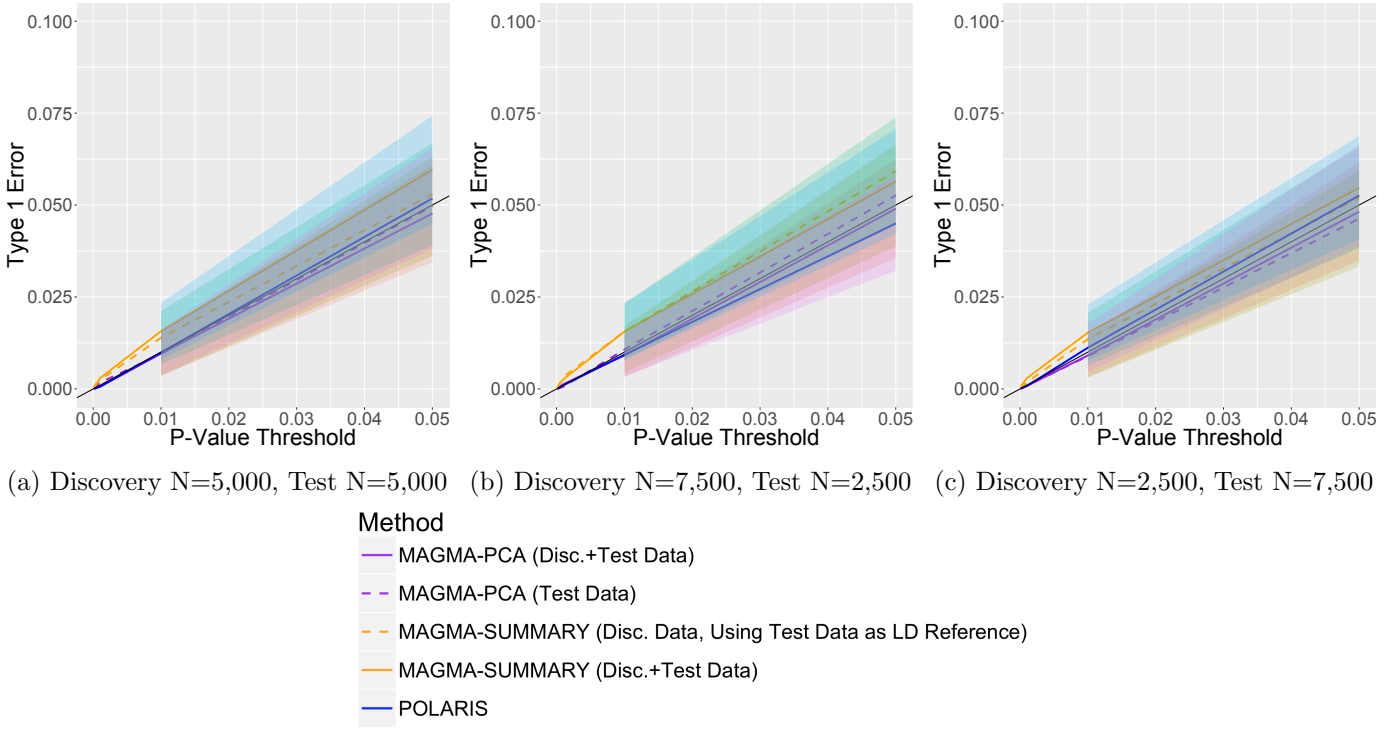


(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 6.12: Type I Error Comparison of Set-Based Methods; Simulation of 10 SNPs with Varying LD with ORs with Randomly Varying Direction and 90 independent, unassociated SNPs. *Note: y-axis scale is not between 0 and 1*

6.3.1.5 Real Data Simulation

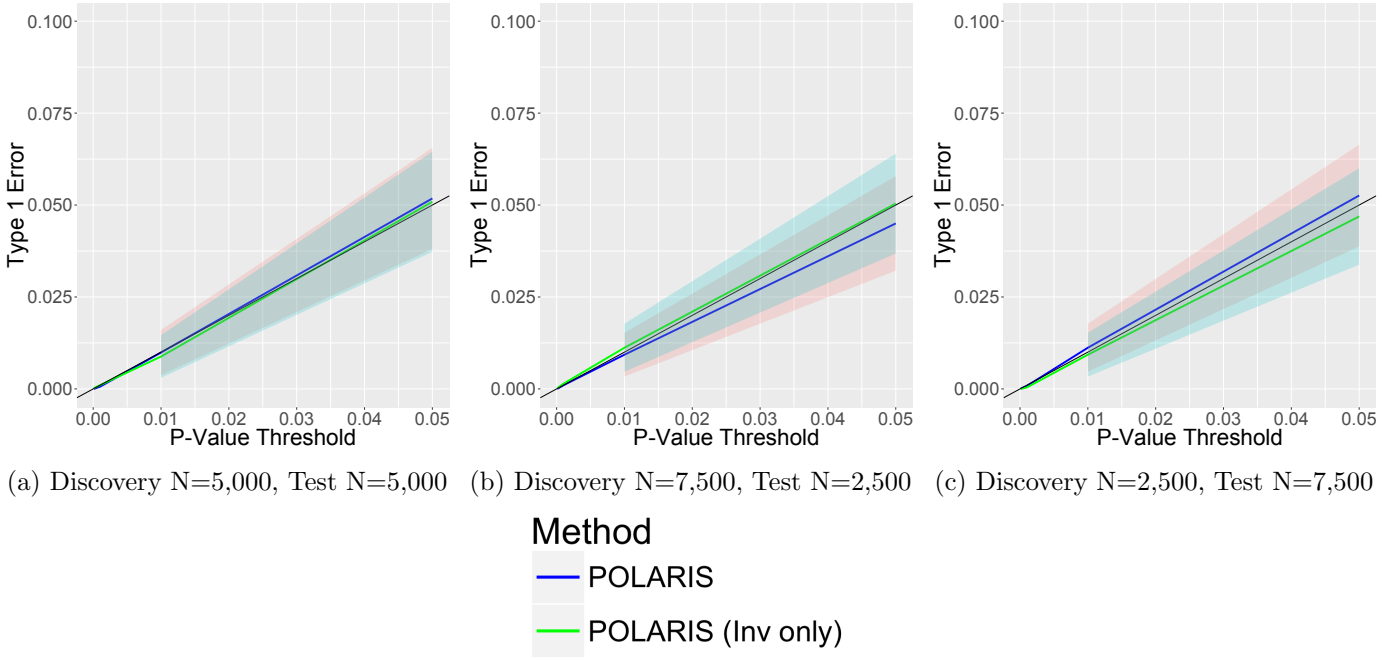
The type I error for the real data simulations, with the case-control status randomly permuted in order to remove the effect size of any SNPs, is shown in Figure 6.13. The LD structure for this scenario can be observed in Figure 6.5. The original (combined) data (N=13164) is split 50/50, 25/75 and 75/25 into independent test and discovery sets. The type I error rate is within 95% CIs of expected values for all data splits.



(a) Discovery N=5,000, Test N=5,000 (b) Discovery N=7,500, Test N=2,500 (c) Discovery N=2,500, Test N=7,500

Figure 6.13: Type I Error Comparison of Set-Based Methods; Simulation of 115 SNPs from Real Data, with Permuted Phenotypes to Remove Effect Sizes. *Note: y-axis scale is not between 0 and 1.*

The comparison between LD adjustment approaches for the real LD structure data is shown in Figure 6.14. The type I error for both adjustment approaches is comparable.



(a) Discovery N=5,000, Test N=5,000 (b) Discovery N=7,500, Test N=2,500 (c) Discovery N=2,500, Test N=7,500

Figure 6.14: Type I Error Comparison of POLARIS and an Adjustment using the Inverse of the Correlation Matrix; Simulation of 115 SNPs from Real Data, with Permuted Phenotypes to Remove Effect Sizes. *Note: y-axis scale is not between 0 and 1.*

6.3.1.6 Power

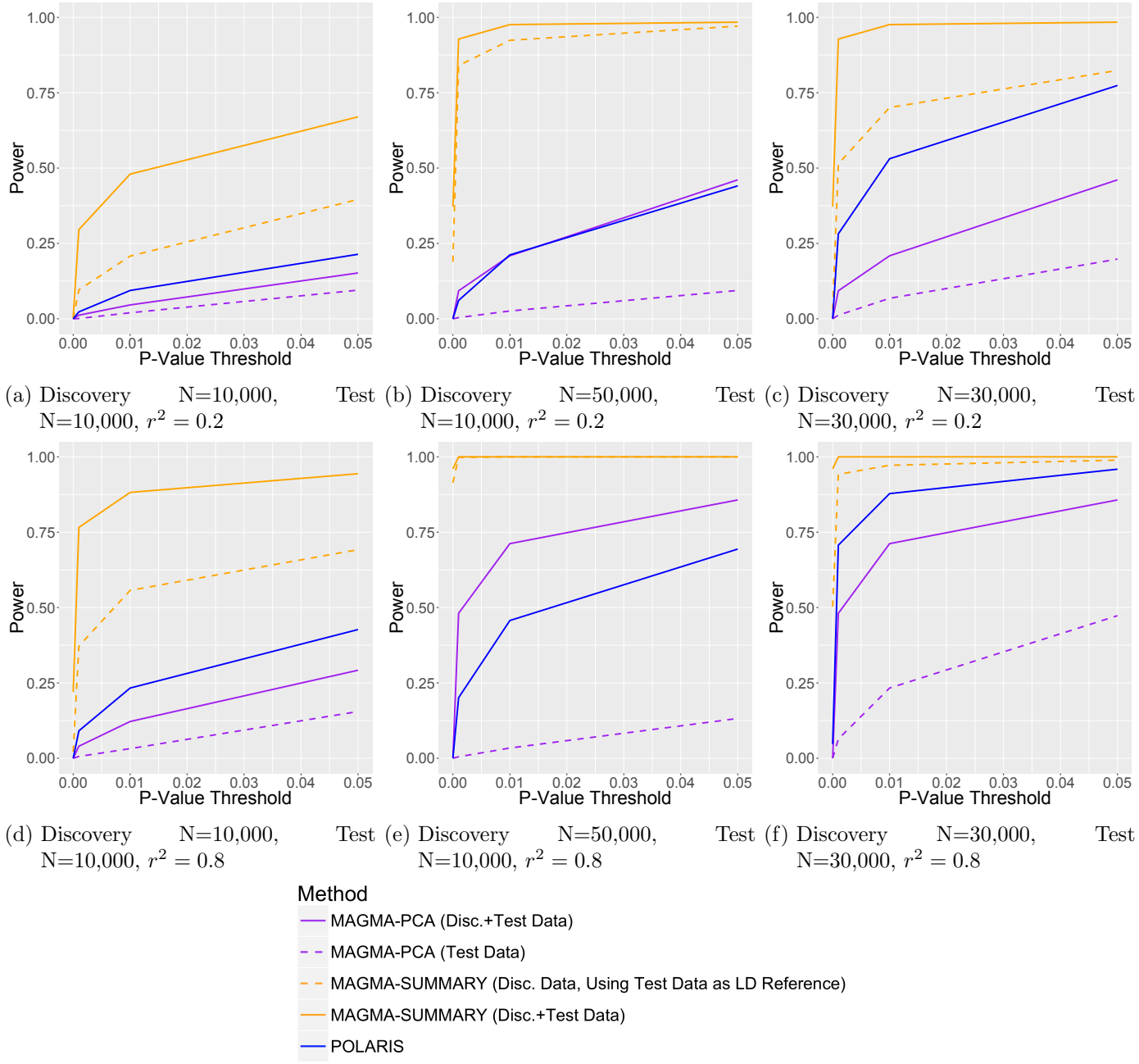
The power of the POLARIS and MAGMA methods are seen in Figures 6.15-6.22. Each Figure displays the simulation results for each scenario discussed in Section 6.2.2. On each graph the solid blue line shows the POLARIS results, the purple dashed line is MAGMA-PCA in the test set only, the purple solid line is MAGMA-PCA in the combined test and discovery sets, the solid orange line is MAGMA-SUMMARY in the combined test and discovery sets and the dashed orange line is MAGMA-SUMMARY in the discovery set using the test set as an LD reference.

6.3.1.7 Simple LD Structure

The power graphs for the simple LD structure are seen in Figure 6.15. The 10 SNPs which are in LD are associated with disease with OR=1.1. POLARIS has equivalent power compared with MAGMA-PCA in the combined discovery and test genotype sets in almost all cases. In the most likely realistic situation, the discovery set is larger than the

test set, but only summary statistics are available for the discovery set. MAGMA-PCA on the combined genotype dataset has higher power where the test N=10,000 and discovery N=50,000, but here MAGMA-PCA is applied to the individual genotypes of the discovery and test sets combined (N=60,000), so the sample used to estimate LD and perform the statistical test is very large, whereas POLARIS uses the discovery set N=50,000 for effect size estimation and only N=10,000 for LD estimation, and importantly, for statistical testing. In all cases, POLARIS has higher power than MAGMA-PCA in the test set only, as is expected, since POLARIS increases power by incorporating information from the discovery set. MAGMA-SUMMARY in both the combined test and discovery summary statistic data and the discovery set data only always has very high power compared to the other methods. It seems unusual that the power is substantially higher using the summary statistics from identical genotype data, and in particular, higher power when considering only the discovery set compared to the combined test and discovery genotype data. Although this effect was demonstrated in simulated data in Section 3.3.3.

The power for POLARIS increases when the size of the test set increases, as this improves the estimate of LD between markers.



(a) Discovery N=10,000, Test N=10,000, $r^2 = 0.2$

(b) Discovery N=50,000, Test N=10,000, $r^2 = 0.2$

(c) Discovery N=30,000, Test N=30,000, $r^2 = 0.2$

(d) Discovery N=10,000, Test N=10,000, $r^2 = 0.8$

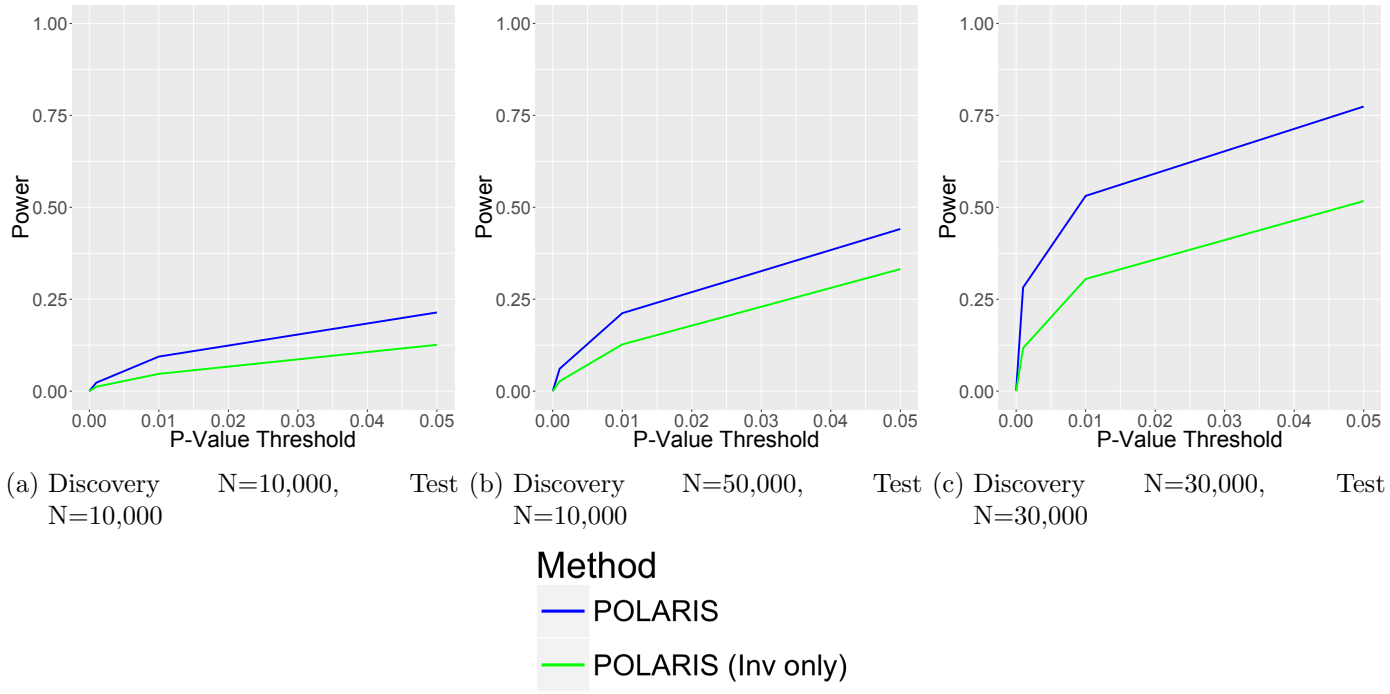
(e) Discovery N=50,000, Test N=10,000, $r^2 = 0.8$

(f) Discovery N=30,000, Test N=30,000, $r^2 = 0.8$

Figure 6.15: Power Comparison of Set-Based Methods; Simulation of 10 SNPs in LD and 90 independent SNPs. Figures 6.15a, 6.15b and 6.15c show Simple LD Structure Simulations where $r^2 = 0.2$, and Figures 6.15d, 6.15e and 6.15f show Simple LD Structure Simulations where $r^2 = 0.8$. Figures 6.15a and 6.15d have a discovery and test sample size of 10,000, Figures 6.15b and 6.15e have a discovery set N=50,000 and test set N=10,000 and Figures 6.15c and 6.15f have discovery and test sets with N=30,000.

Figure 6.16 shows the power comparison between POLARIS which adjusts for LD using the square inverse of the correlation matrix and the LD adjustment using the inverse of the correlation matrix. It is shown that the power using the standard POLARIS LD

adjustment is higher compared to the LD adjustment using the inverse of the correlation matrix only. This is likely because by taking the square inverse rather than the inverse, the adjustment factor is smaller (since $\sqrt{1/\lambda_k} < 1/\lambda_k$), although type I error graphs show that this adjustment is still sufficient.



(a) Discovery N=10,000, Test N=10,000

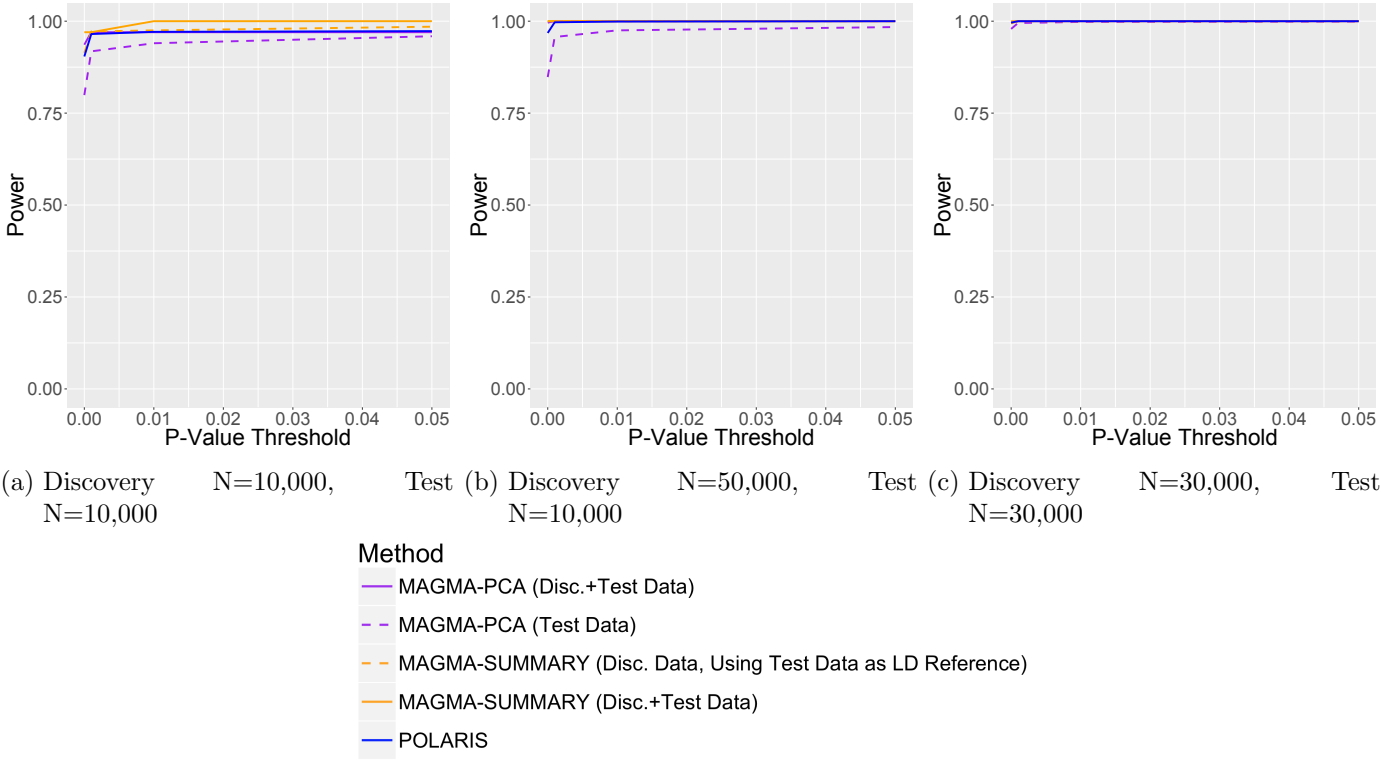
(b) Discovery N=50,000, Test N=10,000

(c) Discovery N=30,000, Test N=30,000

Figure 6.16: Power Comparison of POLARIS and an Adjustment using the Inverse of the Correlation Matrix; Simulation of 10 SNPs in LD and 90 independent SNPs.

6.3.1.8 Complex LD Structure

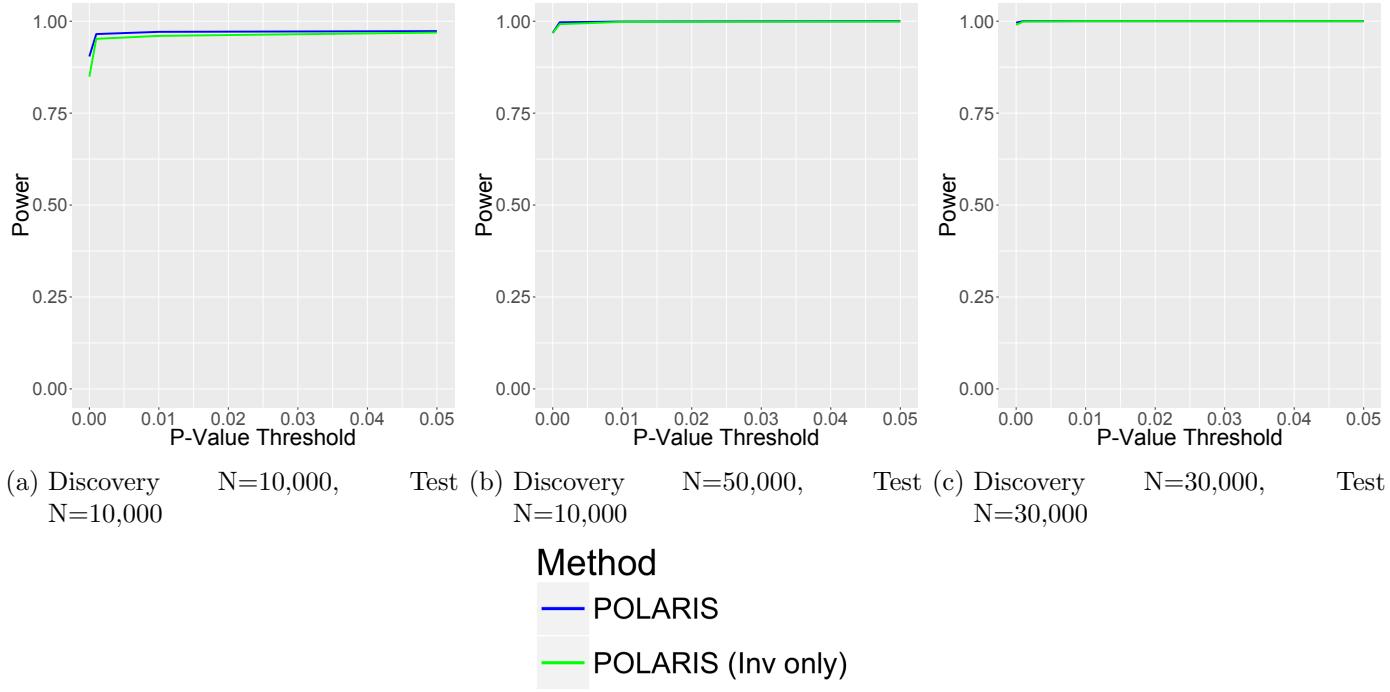
Figure 6.17 shows the power for all methods in the complex LD structure simulations. SNPs in LD were given effect sizes from the following distribution, $OR \sim N(1.02, 0.36)$. All methods have equivalently high power in this simulated data. It is difficult to differentiate between the different methods and conclude which method performs best in this case. Initially, the effect sizes for SNPs in LD were taken from a distribution with mean 1.1, but this was reduced in an attempt to differentiate between the methods. However, because 40 (of the 100 total) SNPs have some association with disease, when combined, the aggregation of these small SNP effects leads to the set of SNPs having a large effect on disease, which explains the high power despite the small individual SNP effect sizes.



(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 6.17: Power Comparison of Set-Based Methods; Simulation of 10 SNPs in LD with $r^2 = 0.2$, 10 SNPs in LD with $r^2 = 0.4$, 10 SNPs in LD with $r^2 = 0.6$, 10 SNPs in LD with $r^2 = 0.8$ and 60 independent, unassociated SNPs.

The power for the two different PRS LD adjustment approaches are seen in Figure 6.18. Again, the power is very high for both approaches and it is therefore difficult to determine which has optimal power in this case. The effect sizes of SNPs was reduced in order to attempt to better differentiate between methods, however, due to the number of SNPs which are associated with disease, this did not help. It may be possible to determine a difference by reducing the overall number of SNPs which are associated with disease.



(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

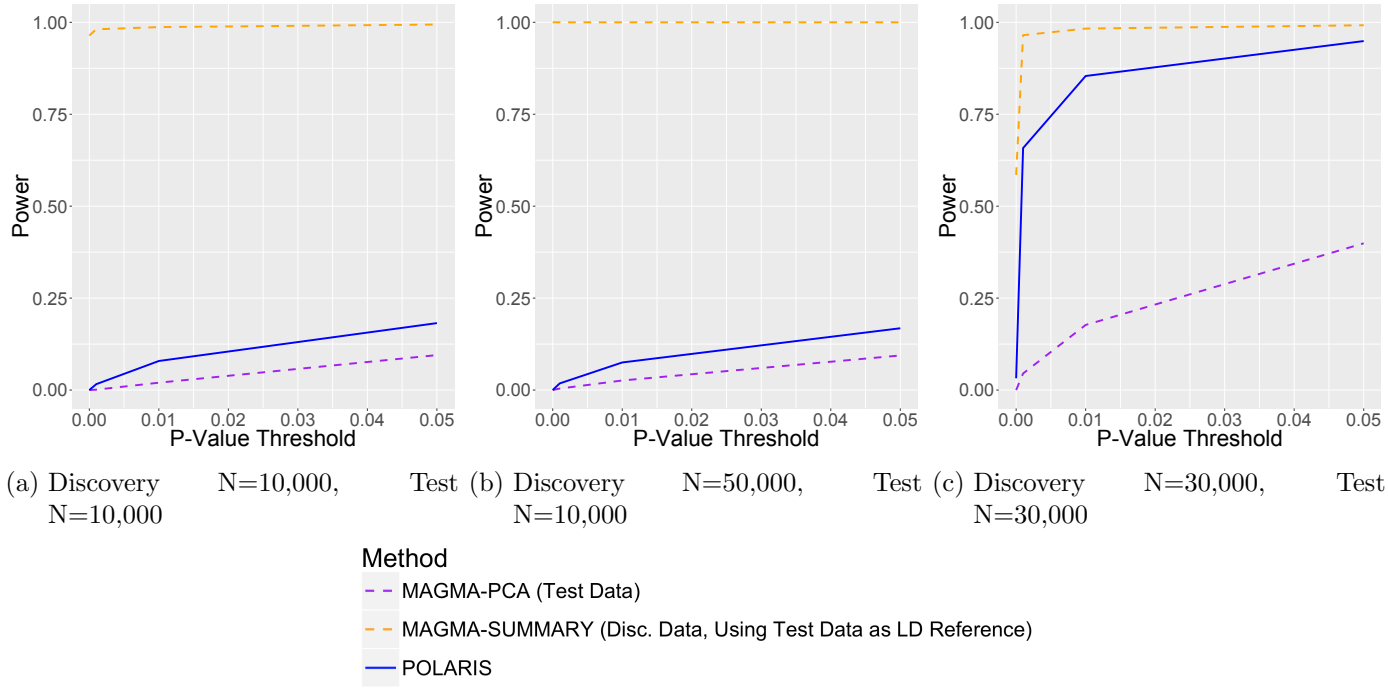
Figure 6.18: Power Comparison of POLARIS and an Adjustment using the Inverse of the Correlation Matrix; Simulation of 10 SNPs in LD with $r^2 = 0.2$, 10 SNPs in LD with $r^2 = 0.4$, 10 SNPs in LD with $r^2 = 0.6$, 10 SNPs in LD with $r^2 = 0.8$ and 60 independent, unassociated SNPs.

6.3.1.9 Different LD Structure of Discovery and Test Datasets

For the simulations where the strength of LD differs between the test and discovery sets, the power graphs are displayed in Figure 6.19. Here, the methods using the test datasets only and the combined sets are not compared; since the combination of two sets with differing LD will result in a set with one average LD strength. This is, of course, different to the LD structure in the test dataset only.

POLARIS is able to increase power using a dataset which has a slightly different LD structure compared to MAGMA-PCA in the test set only. It is usually expected that the LD structures in the test and discovery set will be similar, however, POLARIS is able to utilise this additional data whilst maintaining a self-contained test of association in the test data only. MAGMA-SUMMARY in the discovery set only has very high power in this case, this is most likely caused by the highly inflated type I error. The power of POLARIS increases when the test set sample size increases (N=30,000), this is because the additional subjects in the test set add power to the logistic regression model used to

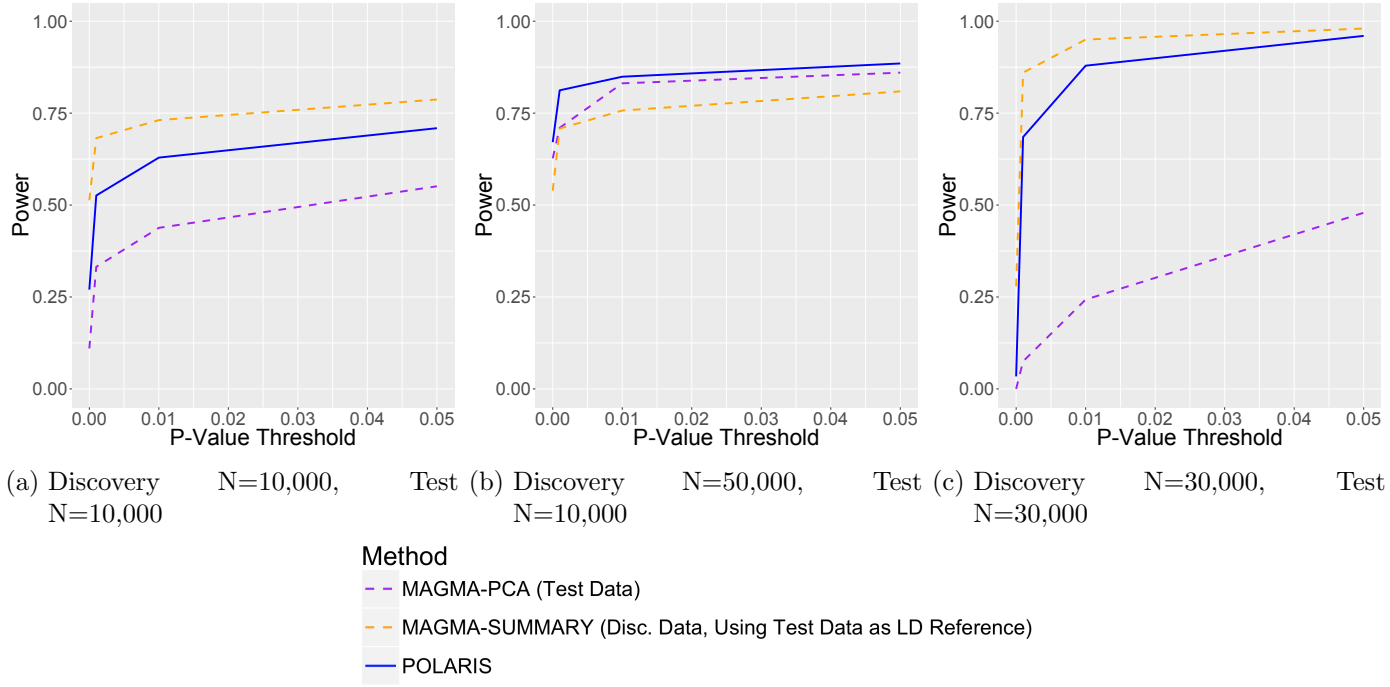
determine the association of the set to disease.



(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 6.19: Power Comparison of Set-Based Methods; Simulation of 10 SNPs in LD with OR$\sim$ N(1.02, 0.2^2) and 90 independent, unassociated SNPs where Test LD is moderate ($r^2 = 0.6$) and Discovery LD is high ($r^2 = 0.8$).

The opposite case where the LD in the test set is high ($r^2 = 0.8$) and LD in the discovery set is moderate ($r^2 = 0.6$) is shown in Figure 6.20. In this case, the MAGMA-SUMMARY method does not have as high power compared to the other methods. This is because the LD will be overadjusted for in this case, rather than underadjusted in the previous case. POLARIS and MAGMA-SUMMARY have similar power, and MAGMA-PCA has lower power compared to POLARIS.

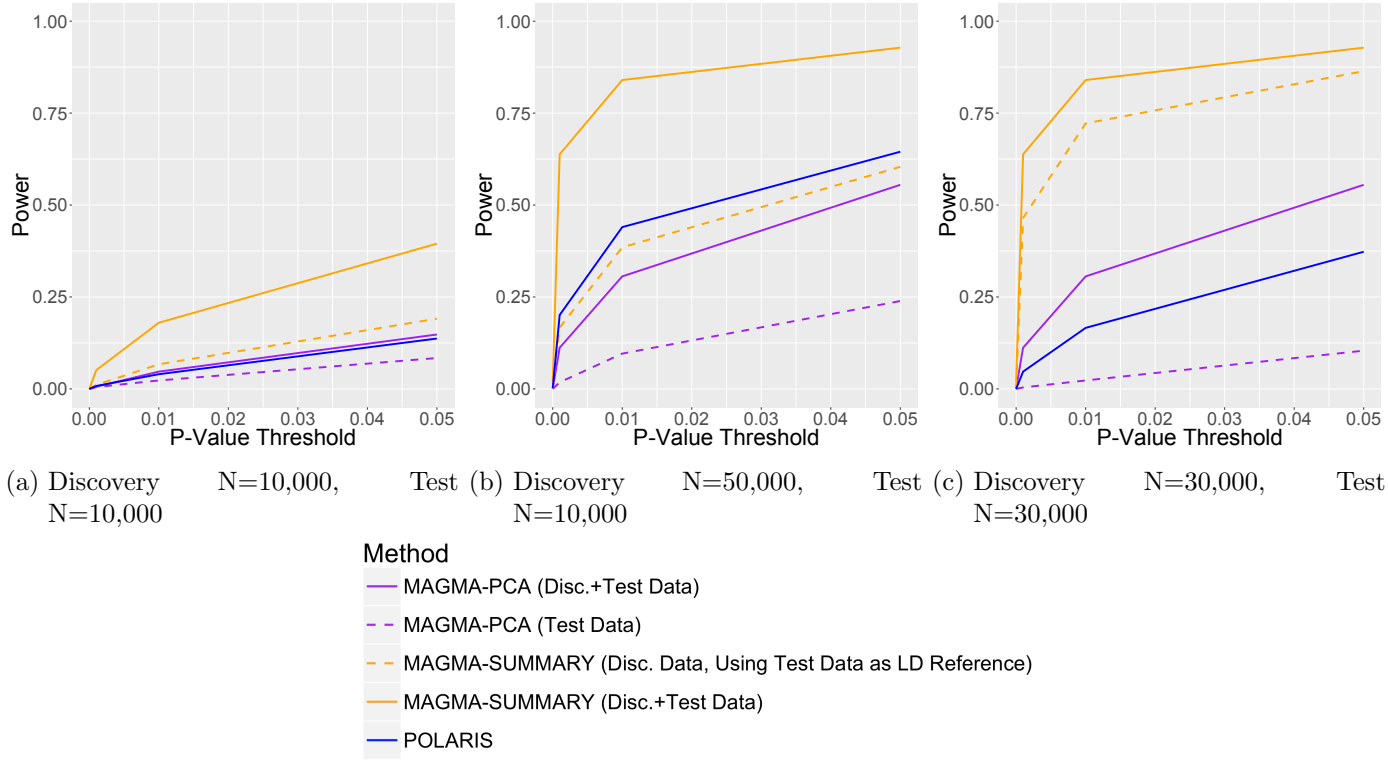


(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 6.20: Power Comparison of Set-Based Methods; Simulation of 10 SNPs in LD with OR$\sim$ N(1.02, 0.2^2) and 90 independent, unassociated SNPs where Test LD is high ($r^2 = 0.8$) and Discovery LD is moderate ($r^2 = 0.6$).

6.3.1.10 Effect Sizes with Varying Direction

Figure 6.21 shows the power for all methods when the SNPs in LD are not always associated with disease in the same direction, both within each separate dataset and across the test and discovery sets. MAGMA-SUMMARY considers p-values only, rather than the direction of effect whereas POLARIS and MAGMA-PCA take account of the direction of effect. MAGMA-SUMMARY in the combined test and discovery sets has highest power compared to the other methods. POLARIS has higher power compared with MAGMA-PCA in the test datasets. POLARIS also has higher power compared to MAGMA-PCA in the combined test and discovery sets when the test dataset has sample size N=10,000. If the researcher is interested in just the set association (independent of direction) then MAGMA-SUMMARY is appropriate, however, if directionality is important, then POLARIS or MAGMA-PCA should be used.

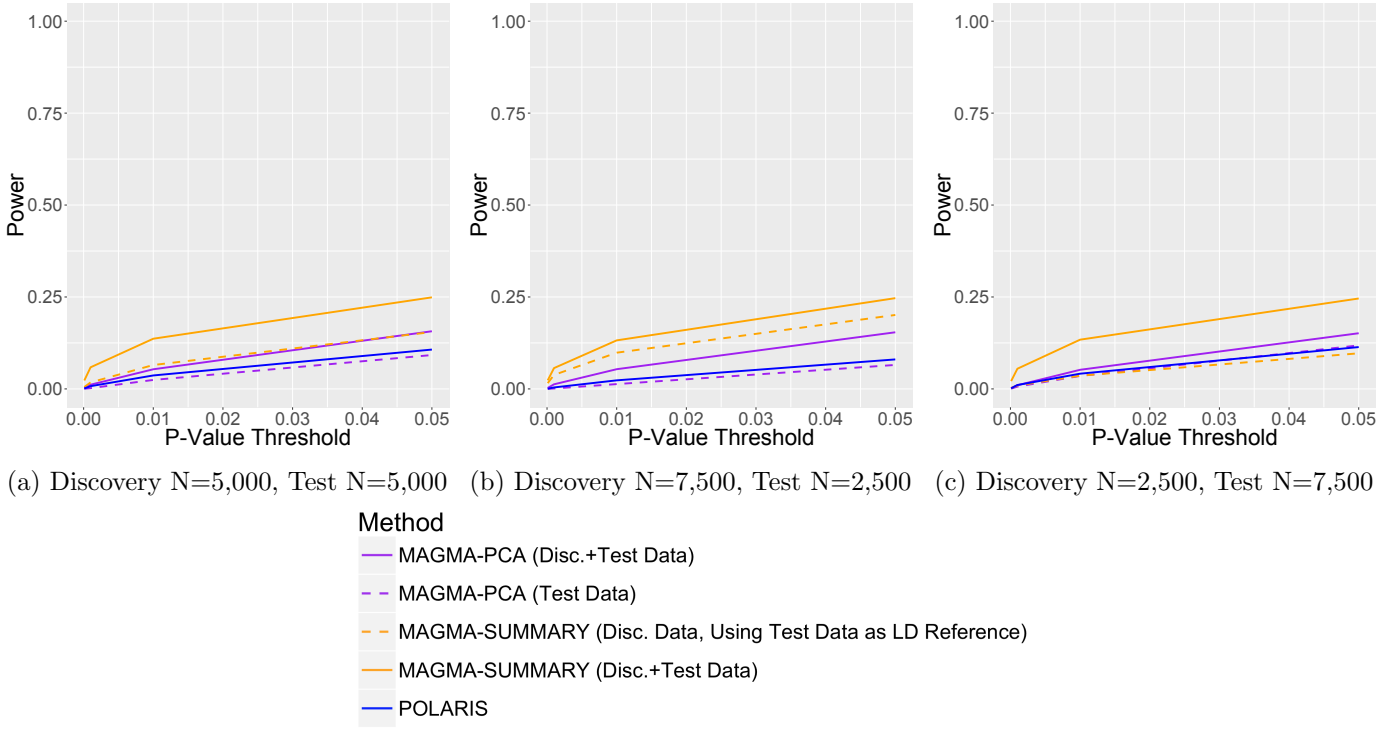


(a) Discovery N=10,000, Test N=10,000 (b) Discovery N=50,000, Test N=10,000 (c) Discovery N=30,000, Test N=30,000

Figure 6.21: Power Comparison of Set-Based Methods; Simulation of 10 SNPs with Varying LD with ORs with Randomly Varying Direction and 90 independent, unassociated SNPs.

6.3.1.11 Real Data Simulation

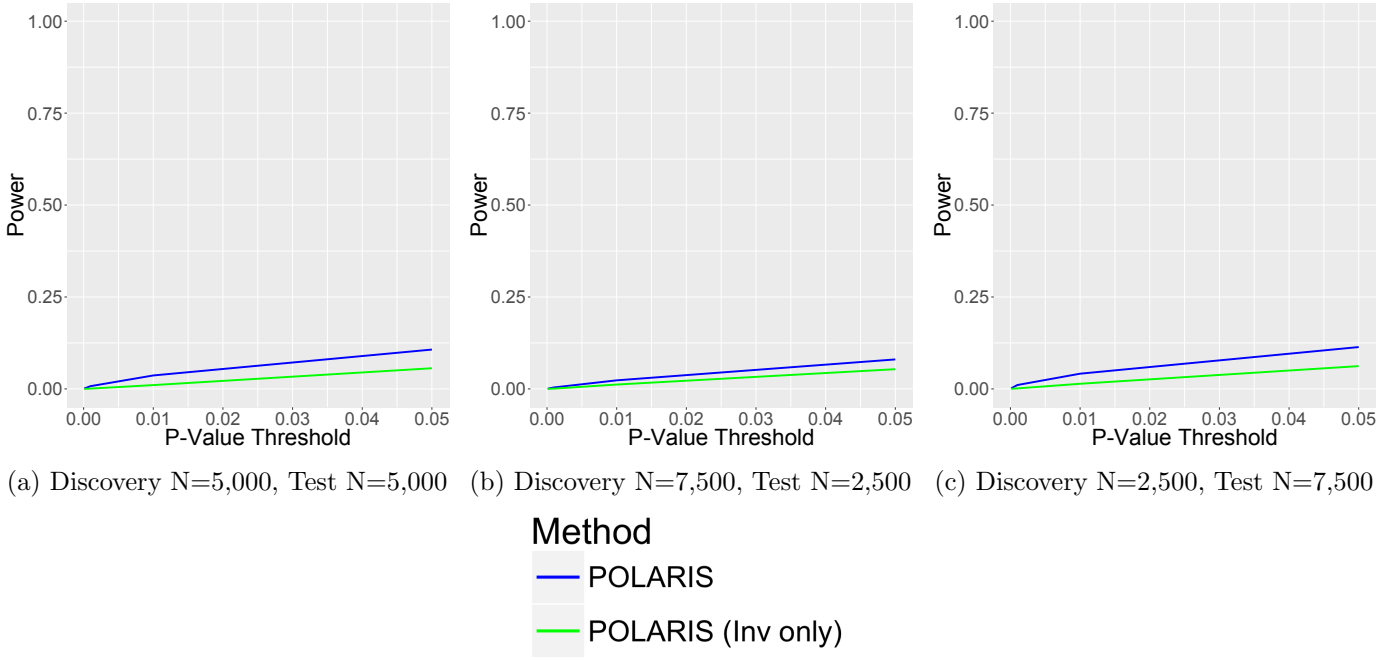
Figure 6.22 shows the power graph for the real data simulations. The power of the POLARIS method lies generally between the power of MAGMA-PCA applied to the test set only and MAGMA-PCA in the combined test and discovery sets. One can see that by using the information from the discovery set, POLARIS increases the power compared to using the test set only, but, as is to be expected, not as much as using the individual genotypes from the discovery set as well as the test set. This increase seen in POLARIS is reduced when the test set has a large number of individuals (N=7,500) relative to the discovery set (N=2,500) (see Figure 6.22c). Also note that the power of the MAGMA-SUMMARY approach exceeds the power of MAGMA-PCA on the same combined dataset, a demonstration of this effect is discussed in Section 3.3.3.



(a) Discovery N=5,000, Test N=5,000 (b) Discovery N=7,500, Test N=2,500 (c) Discovery N=2,500, Test N=7,500

Figure 6.22: Power Comparison of Set-Based Methods; Simulation of 115 SNPs from Real Data

Figure 6.23 shows the power comparison between the standard POLARIS LD adjustment using the square inverse of the correlation matrix and the LD adjustment using the inverse of the correlation matrix only. The power for the standard POLARIS approach has higher power in the simulated data with a real LD structure, although this difference is very slight, due to the relatively small power for both approaches.



(a) Discovery N=5,000, Test N=5,000 (b) Discovery N=7,500, Test N=2,500 (c) Discovery N=2,500, Test N=7,500

Figure 6.23: Power Comparison of POLARIS and an Adjustment using the Inverse of the Correlation Matrix; Simulation of 115 SNPs from Real Data

6.4 Discussion

The main aim of this chapter was to develop a novel methodology which accounts for LD in the calculation of a PRS. The resulting individual LD-adjusted PRS can also be used for analysing whether a set of SNPs is associated with disease. This method combines the advantages of PRS and spectral analysis of the genetic data. The latter suggests a mathematically sound adjustment for LD and includes a stabilisation parameter (similar to ridge regression) to cope with cases of extreme LD. It adjusts for LD between SNPs and informs the analysis with previously reported effect sizes of a SNP's association with disease. This adjustment uses the SNP-SNP correlation matrix; however, one could alternatively use the SNP-SNP covariance matrix. For all examples above, this gives very similar results. The benefits of using the correlation matrix over the covariance matrix are that SNPs with low MAF are not penalised, i.e. their contribution to the risk score is equivalent to that of a more common SNP. When partitioning the overall polygenic risk based on meaningful SNP sets, the method allows both to test for significance of association of these sets (set-based analysis) and to provide individual set-specific risk scores for subjects, which can be further used for risk prediction of subphenotypes with respect to

the SNP sets.

To assess the quality of the proposed approach, POLARIS for set-based analysis was compared with the widely used MAGMA software. POLARIS was shown to give the correct type I error and its power lies between that of MAGMA-PCA applied to the test dataset only and MAGMA-PCA applied to the combined test and discovery datasets. MAGMA-SUMMARY is shown to have higher power compared to MAGMA-PCA often in identical data. In practice, researchers would use all the available genotype data, and would use PRS based methods if effect sizes only are known for an additional dataset.

The LD adjustment in POLARIS uses the square inverse of the correlation matrix. This was compared to the use of the inverse of the correlation matrix. The type I error was consistent for both LD adjustment approaches and the power is at least equivalent and often higher for the standard POLARIS LD adjustment.

POLARIS has three main advantages. 1) It produces a risk score per person per set, unlike other set-based methods which only provide a p-value for the strength of association between the set and disease. This set risk score can be used to stratify individuals for follow up studies (e.g. clinical trials) and also prioritise genes for further functional studies (e.g. animal models), supporting the development of precision medicines. 2) POLARIS can increase power by leveraging the discovery set to perform a self-contained test of association in the test dataset. Another way to incorporate the discovery set would be to use meta-analysis, however, this detects an association in the combined set rather than the test set only. This may be important when the test data differs in some way from the discovery data, e.g. different ethnicity, or different phenotype. A good example might be where the test sample uses different diagnostic criteria to measure the same phenotype (e.g. self-report questionnaire for depression) and one wishes to validate these criteria by showing that they show association to the same genes as those implicated by the standard diagnosis. 3) The overall set association can easily be adjusted by population or any other covariates.

Like the $PRS_{set-based}$ method, POLARIS can be used for any set of SNPs, for example the whole genome, genes or pathways. Therefore it has the potential of data driven discovery

of pathways.

POLARIS can also be utilised in a number of cross disorder analyses to determine commonality between disorders at a gene-based or pathway-based level. There are a number of common disorders for which the GWAS summary data are publically available (e.g. Psychiatric Genomics Consortium). The GWAS data for one disorder can be used to generate scores per person per gene in another disorder or subphenotype of interest, and thus test for overlap between disorders at a gene-based level.

The POLARIS method can be extended to add additional information into the score, such as rare variants from exome sequencing studies. The POLARIS set-based method is implemented into a freely accessible platform independent software, see Supplementary section 11.3 for scripts. Large sets have a high computational burden due to the spectral decomposition of large correlation matrices; recommendations on maximum set size and the corresponding required computational resource are included with the software.

In this study, POLARIS was applied to test binary traits. However, the POLARIS score can also be used as a variable (along with other covariates) in regression models for quantitative traits.

A limitation of the POLARIS implementation is that currently it is only available as a self-contained set-based method. However, POLARIS can in principle be used as a competitive set analysis, adjusting for the baseline level of association in the data either by including a general PRS in the analysis or comparing the set-based PRS to those generated from random sets of genes (matched for number of genes/gene size/numbers of SNPs) or random sets of SNPs (matched for LD, MAF and SNP density).

Another limitation of PRS-type approaches is the imperfect tagging of the underlying causal variants by SNPs and imperfect effect size estimates. The challenge of selecting the true set of susceptibility SNPs for PRS modelling to capture heritability has been pointed out [99]. Our approach can use all SNPs in a set of interest, even when in LD, and therefore any causal genotyped SNPs will be included. If the causal SNPs are not present in the sample, then the tagging SNPs only are used. The effect sizes of all SNPs in

LD will be adjusted according to the LD structure, not according to the causal/non-causal nature of the SNP.

POLARIS is a valuable extension to standard PRS, by adjusting for LD between markers and removing the necessity to LD prune data prior to analysis. POLARIS provides a test of the set's association with disease whilst also producing subject specific risk scores.

7 POLARIS: Gene-Based and Pathway Analyses in AD Data

7.1 Introduction

Set-based analysis is an alternative to single SNP analyses, and may be more powerful due to the aggregate effect of multiple SNPs being larger than that of individual SNPs. For example, determining the association of genes rather than SNPs is beneficial since genes are more robust across different populations [44]. Set-based analyses are being widely used in the literature and as expected, are able to identify novel genes or pathways associated with disease. Gene-sets, or pathways, have been assessed and eight have been found to be associated with AD using the ALIGATOR [61] algorithm [23]. Additionally 134 gene-sets have been identified as being associated with SZ which are related to nervous system function and development, where gene-sets are defined from single gene functional studies [100]. Other gene and gene-set analyses were considered in a SZ study investigating exomic variation which determined an enrichment in genes whose Messenger Ribonucleic Acid (mRNA) binds to Fragile X Mental Retardation Protein (FMRP) and LoF intolerant genes [101].

As illustration, the power increase from using a gene-based analysis is shown using MAGMA, see Chapter 3, where additional genes not discovered using single SNP analyses are determined. In order to further improve power, PRS was applied to the set-based framework, see Chapter 4, this incorporates additional information from external data into the analysis and in addition, provides a risk score per subject for each set. In fact, the use of

PRS gene-based method in AD imputed data in Chapter 5 led to the identification of two novel genes: *CSMD1* and *MACROD2*. However, this standard PRS approach requires independence between SNPs, and therefore the data must be pruned for LD prior to the analysis.

A pathway analysis offers a higher level of information aggregation, by integrating multiple genes into a pathway, which implicates a particular biological process. A pathway analysis has been applied to the AD IGAP data using ALIGATOR [61] and eight pathways have been found to be associated with AD; these are immune response, regulation of endocytosis, cholesterol transport, hematopoietic cell lineage, proteasome ubiquitin activity, reactome hemostatis, clathrin and protein folding [23][24]. The PRS set-based approach was applied to these eight pathways, see Chapter 5, and two pathways were consistently associated using this method; these were the immune response and hematopoietic cell lineage pathways. The correlation among these eight pathways was also high ($p < 2.2 \times 10^{-16}$), although the actual correlation coefficient was often low (r ranges from 0.0201 to 0.3385). This correlation suggests that some SNPs are shared between different pathways.

To remove the issue with the PRS set-based method where LD pruning is required, POLARIS was introduced, see Chapter 6. POLARIS extends PRS by incorporating an LD adjustment between SNPs. POLARIS was shown to have adequate power which resided between MAGMA-PCA in the test set only and MAGMA-PCA in the combined test and discovery sets. In this chapter POLARIS is applied to real AD data, similarly to Chapter 5. The aim of this chapter is to investigate the power of POLARIS when applied to real AD data and compare the results with MAGMA-based analyses.

A number of set-based analyses are known to introduce bias in the set-based p-value due to the size of the set [84][95]. SNP p-values may contain a small amount of inflation, possibly due to population stratification. When a large number of these SNPs are combined, this inflation can become substantial, therefore, it was assessed whether POLARIS is influenced by this bias.

Previous studies [102] have identified that genes found from gene-based analyses are evolutionary constrained, likely due to functional importance. Genes found from the POLARIS

gene-based analysis were interrogated to determine if they reside in CNS or whether there is an enrichment of genes which are LoF intolerant. It is expected that no enrichment will be found, since AD is a post-reproductive disorder.

POLARIS can be implemented in either directly genotyped or imputed data. Imputation is when the genotypes of a set of SNPs is inferred. This is done using the LD between untyped SNPs and typed SNPs from a reference panel [103]. A number of different softwares are available for imputation, such as BEAGLE [104], MaCH [105] and IMPUTE2 [106]. Data which has been directly genotyped is less likely to contain incorrect genotypes, and is quick to run computationally, due to the smaller number of SNPs. However, the power of the analyses can be increased by using imputed data due to the inclusion of a larger number of SNPs. The imputation accuracy can be controlled by info score or the probability of correct imputation. These are information metrics which typically take a value between 0 and 1, with 1 indicating high certainty imputation for a SNP; a cutoff for these metrics of a value greater than 0.4 is often used to remove SNPs which have been imputed with poor accuracy [103].

7.1.1 Objectives

This Chapter focuses on using POLARIS applied to real AD data by aiming to:

- Perform a gene-based analysis using POLARIS in the GERAD genotype data.
- Compare POLARIS real data gene-based results with those found using MAGMA-PCA and MAGMA-SUMMARY.
- Perform POLARIS gene-based analysis in the GERAD imputed data.
- Compare the POLARIS gene-based results using genotyped and imputed data.
- Investigate whether POLARIS is biased by the number of SNPs in the gene and compare this to MAGMA approaches.
- Demonstrate whether genes from the gene-based analysis are represented in con-

served regions.

- Compute PRS pathway scores for the eight pathways previously found to be associated with AD [23][24] and determine whether these pathways are associated with AD when tested for association using the POLARIS approach.

- Determine whether any correlation is observed between the different POLARIS pathways to assess whether there is any statistical commonality between pathways.

7.2 Materials and Methods

POLARIS was applied to the largest and most powerful AD dataset (see detailed description in Section 2.1). The GERAD data was used as the test set (3,332 cases, 9,832 controls), and IGAP data (17,008 cases, 37,154 controls) excluding GERAD subjects was used as the discovery set, and is thus used to improve power. A set-based risk score was produced for every individual in the GERAD data.

POLARIS adjusts for LD between SNPs and therefore, the SNPs were not pruned for LD and the entire data were used in this analysis. There were 419,048 SNPs in common between the GERAD and IGAP data.

It was necessary to ensure that SNP alleles were coded in the same direction across both the discovery (IGAP excluding GERAD subjects) and test (GERAD) datasets. This is because summary statistics provide an effect size with respect to a particular allele and this allele must be the same in both datasets. If alleles in IGAP-noGERAD were coded in the opposite direction to those in GERAD, the summary effect size for the SNP was inverted. SNPs with alleles AT, TA, CG or GC were excluded since the direction of the effect could not always be determined when combining two studies. Of the SNPs in IGAP-noGERAD, 103,356 matched those in GERAD, the remaining had effect sizes inverted and no SNPs were excluded due to ambiguity. An MAF filter of 0.01 was applied to the data.

POLARIS was also implemented to process the data as imputed with HRC, Chapter 5 also considered the use of gene-based methods in both genotype and imputed data. Again, the

complete data were used, since LD pruning was not required, a total of 3,169,840 SNPs were in common between imputed GERAD and IGAP data.

7.2.1 POLARIS Gene-Based Analysis

POLARIS, MAGMA-PCA and MAGMA-SUMMARY were applied to the AD data to determine gene-wide p-values in two cases. For the first case, only directly genotyped SNPs from the GERAD data were used (cf. [21], where however, imputed genotype data were used for IGAP summary statistics analysis). The second case considers the HRC imputed GERAD data.

SNPs were assigned to genes using GENCODE (v19) gene models [78]. Only genes with known gene status and those marked as protein coding were used. Two different gene windows were considered, the first used no window around the gene, only SNPs within the start and end position of the gene were included, and the second considered SNPs which were within 35kb upstream and 10kb downstream of the gene. SNPs which belong to multiple genes were assigned to all those genes.

For the genotype GERAD data, a total of 202,504 SNPs were assigned to 14,620 distinct genes with a maximum of 1,342 SNPs in a gene and for the HRC imputed GERAD data, 1,122,570 SNPs were assigned to 17,072 distinct genes.

The missing genotypes in real data were imputed as in PLINK [50][51], where missing genotypes are substituted by $2 \times MAF$ for each SNP. In the GERAD data, 0.0514% of genotypes required imputation.

The results of gene-based analyses for AD data using POLARIS were compared to those from the MAGMA-PCA approach in GERAD data and also the MAGMA-SUMMARY approach in IGAP data (it was not possible to use MAGMA-PCA in the full IGAP sample, as the individual genotypes were not available). For the latter, only SNPs present in both IGAP and GERAD are considered. Prior to the gene-based analysis, SNP summary statistics for the whole IGAP data were adjusted for the genomic control parameter, λ=1.087, as reported in [21]. Many gene-based methods which use summary statistics are

inflated by the number of SNPs in a gene, this is due to the aggregration of small errors in single SNP effect size estimation [79][95].

It was then assessed whether genes determined from the gene-based analysis were enriched in conserved regions; both for genes which are evolutionary constrained and those which reside in CNS. These regions are less likely to harbour variants of a strong effect or are less prone to variation. The number of genes from the gene-based analysis with a p-value below either a nominal (0.05) or gene-wide threshold (2.5×10^{-6}) and were in conserved regions were determined. The 2x2 contingency tables have the number of genes with p-values above and below the p-value threshold and the rows show the number of genes in/out of LoF regions. The strength of association between location in LoF regions and p-value was assessed using a chi-squared test. When the cell counts in the 2x2 table were small a Fisher's exact test was used to determine the association between gene significance and location in LoF regions.

7.2.2 POLARIS Pathway Analysis

Pathway analyses can provide insights and lead to biological mechanisms of disease.

A POLARIS score was produced for every GERAD subject for each of the eight pathways found to be associated with AD (immune response, regulation of endocytosis, cholesterol transport, hematopoietic cell lineage, proteasome ubiquitin activity, reactome hemostasis, clathrin and protein folding) [23][24]. A self-contained and competitive test of association was performed for each of these pathway scores. A self-contained test was implemented using a logistic regression model of the POLARIS pathway scores regressed with the AD phenotype. A competitive test for each pathway was performed using a likelihood ratio test to find the additional benefit of including the POLARIS pathway score to a model including POLARIS scores across the whole genome.

Only SNPs with a p-value less than 0.5 in the IGAP study are included into the pathway POLARIS risk score, since it was shown that maximum prediction using PRS was attained using a p-value threshold of 0.5 [38]. The pathway risk score was used to find

the association strength of the pathway by including the risk score in a logistic regression model which adjusts for other population covariates.

The pairwise correlation between all eight POLARIS pathways was also investigated in order to determine any commonality between the pathways. The Pearson correlation coefficient was used (cor.test() in R) to test for all pairwise correlations between all eight pathways.

7.3 Results

7.3.1 POLARIS Gene-Based Analysis

Table 7.1 demonstrates the number and proportion of genes below a particular p-value threshold for POLARIS (combining IGAP-noGERAD summary statistic data and GERAD), MAGMA-PCA in GERAD genotype data and MAGMA-SUMMARY in IGAP summary statistic data. The table shows that there are many more significant associations for MAGMA-SUMMARY as compared to POLARIS and MAGMA-PCA. However, these results may be misleading as statistically independent associations in some instances implicate overlapping regions. To define genes as physically independent, we have annealed associated genes that were not separated by at least 250kb in each analysis separately and have taken the most associated gene. In the *APOE* region, significant genes on chromosome 19 between 44.4Megabase (Mb) and 46.5Mb were counted as one. The results for the independent best genes are presented in Table 7.2.

The number of independent significant genes for all p-value thresholds is higher or equal for POLARIS compared to the MAGMA-PCA approach in GERAD data. This is expected as POLARIS uses both GERAD and IGAP-noGERAD data, while MAGMA-PCA uses GERAD genotypes only. The results for the summary statistic approach show higher numbers of significant genes for higher significance thresholds. The five gene-wide significant genes found by the summary statistics approach are: *TOMM40*, *CLU*, *BIN1*, *MS4A4E* and *CR1*, which have all been previously reported as being associated with AD from single

SNP analyses [19] [20]. For these five genes, POLARIS also finds an association, but does not always reach gene-wide significance ($p = 6.33 \times 10^{-24}$, $p = 7.17 \times 10^{-6}$, 0.00112, 0.00108 and 0.00065 respectively). The difference between the results for MAGMA-SUMMARY and MAGMA-PCA in the same data could be explained by the inflation seen in the MAGMA-SUMMARY approach when associated SNPs are in LD, see Section 3.3.3 for further explanation.

Table 7.1: Comparison of the Number and Proportion of All Genes Below a P-value Threshold for POLARIS, MAGMA-PCA in GERAD data and MAGMA-SUMMARY in IGAP data

P-value Threshold	POLARIS		MAGMA-PCA in GERAD		MAGMA-SUMMARY in IGAP	
	No. of Genes	Prop. of Genes	No. of Genes	Prop. of Genes	No. of Genes	Prop. of Genes
1*	14620		14606		14607	
0.05	840	0.0575	749	0.0513	783	0.0536
0.01	192	0.0131	162	0.0111	244	0.0167
0.001	30	0.0021	24	0.0016	62	0.0042
0.0001	13	0.0009	9	0.0006	31	0.0021
0.00001	6	0.0004	4	0.0003	21	0.0014
0.000001	4	0.0003	3	0.0002	15	0.0010

* Note that the total number of genes (p-values threshold equal to 1) differs, this is due to some gene exclusions made by MAGMA software.

Table 7.2: Comparison of the Number and Proportion of **Independent** Genes Below a P-value Threshold for POLARIS, MAGMA-PCA in GERAD data and MAGMA-SUMMARY in IGAP data

P-value Threshold	POLARIS		MAGMA-PCA in GERAD		MAGMA-SUMMARY in IGAP	
	No. of Genes	Prop. of Genes	No. of Genes	Prop. of Genes	No. of Genes	Prop. of Genes
1*	563		581		560	
0.05	302	0.5364	283	0.4871	255	0.4554
0.01	116	0.2060	98	0.1687	114	0.2036
0.001	19	0.0337	12	0.0207	31	0.0554
0.0001	7	0.0124	4	0.0069	12	0.0214
0.00001	3	0.0053	2	0.0034	9	0.0161
0.000001	2	0.0036	1	0.0017	5	0.0089

* Note that the total number of genes (p-values threshold equal to 1) differs, this is due to some gene exclusions made by MAGMA software.

The number and proportion of genes below particular p-value thresholds are seen in Table

7.3 for the POLARIS gene-based analysis, MAGMA-PCA approach in GERAD imputed data and MAGMA-SUMMARY approach in IGAP data containing only SNPs in the GERAD imputed data. The MAGMA-SUMMARY method determines the largest number of genes at the most stringent p-value thresholds compared to POLARIS and MAGMA-PCA. POLARIS finds a larger number of genes at less stringent thresholds ($p > 0.001$) compared to both MAGMA approaches. However, some genes may be in overlapping regions, therefore, Table 7.4 is presented which includes only independent genes. Genes are defined to be independent by annealing associated genes that were not separated by at least 250kb and have taken the most associated gene. For the *APOE* region, all significant genes on chromosome 19 between 44.4Mb and 46.5Mb were counted as one. Considering just independent genes, the difference between the MAGMA-SUMMARY method is less substantial, but a greater number of genes are still determined at low p-value thresholds. This approach finds 4 additional genes compared to POLARIS; *PICALM*, *CR1*, *MS4A4E* and *BIN1*, all of which have previously been identified as being associated with AD. These genes did not attain gene-wide significance in the POLARIS gene-based analysis (p=0.0172, p=0.00146, p=0.0002 and p=0.000821 respectively). The difference between the results using MAGMA-PCA and MAGMA-SUMMARY in the same data is interesting, and is potentially explained by increased power in MAGMA-SUMMARY when there is LD between associated SNPs, see Section 3.3.3 for further details.

Table 7.3: Comparison of the Number and Proportion of All Genes Below a P-value Threshold for POLARIS, MAGMA-PCA in GERAD imputed data and MAGMA-SUMMARY in IGAP data

P-value Threshold	POLARIS		MAGMA-PCA in GERAD		MAGMA-SUMMARY in IGAP	
	No. of Genes	Prop. of Genes	No. of Genes	Prop. of Genes	No. of Genes	Prop. of Genes
1*	17072		14605		14430	
0.05	2037	0.1193	815	0.0558	808	0.0560
0.01	671	0.0393	179	0.0123	282	0.0195
0.001	142	0.0083	21	0.0014	79	0.0055
0.0001	31	0.0018	6	0.0004	46	0.0032
0.00001	9	0.0005	4	0.0003	20	0.0014
0.000001	5	0.0003	3	0.0002	15	0.0010

* Note that the total number of genes (p-values threshold equal to 1) differs, this is due to some gene exclusions made by MAGMA software.

Table 7.4: Comparison of the Number and Proportion of **Independent** Genes Below a P-value Threshold for POLARIS, MAGMA-PCA in GERAD imputed data and MAGMA-SUMMARY in IGAP data

P-value Threshold	POLARIS		MAGMA-PCA in GERAD		MAGMA-SUMMARY in IGAP	
	No. of Genes	Prop. of Genes	No. of Genes	Prop. of Genes	No. of Genes	Prop. of Genes
1*	557		556		556	
0.05	392	0.7038	270	0.4856	251	0.4514
0.01	247	0.4434	112	0.2014	128	0.2302
0.001	92	0.1652	11	0.0198	38	0.0683
0.0001	22	0.0395	2	0.0036	16	0.0288
0.00001	5	0.0090	2	0.0036	8	0.0144
0.000001	2	0.0036	1	0.0018	6	0.0108

* Note that the total number of genes (p-values threshold equal to 1) differs, this is due to some gene exclusions made by MAGMA software.

In the HRC imputed GERAD data, the SNPs are annotated to 17,072 genes. Of these genes, six reach gene-wide significance, compared to the four genes found using the GERAD genotype data only. These genes are seen in Table 7.5. The gene-wide signifcant genes are *CLU*, *BCL3*, *PVRL2*, *TOMM40*, *APOE* and *CLPTM1*; these have all been previously identified as being associated with AD. The majority of these associations are influenced by *APOE* since these genes are close in location to the *APOE* gene. These results are shown on the Manhattan plot in Figure 7.1, the gene-wide significant genes are shown in red, and those with a suggestive p-value ($2.5 \times 10^{-6} < p < 0.00001$) are seen in blue; these are *DAB1*, *ZNF35* and *RORA*.

Table 7.5: Gene-Wide Significant Genes from POLARIS Gene-based Analysis in AD Imputed Data

			POLARIS		
Chr	Gene	No. of SNPs	β	SE	P-value
8	*CLU*	25	0.521	0.1048	6.8×10^{-7}
19	*BCL3*	5	0.927	0.1499	6.1×10^{-10}
19	*PVRL2*	95	0.637	0.0532	5.7×10^{-33}
19	*TOMM40*	20	0.454	0.0351	2.9×10^{-38}
19	*APOE*	1	2.247	0.2679	4.9×10^{-17}
19	*CLPTM1*	93	0.461	0.0965	1.8×10^{-6}

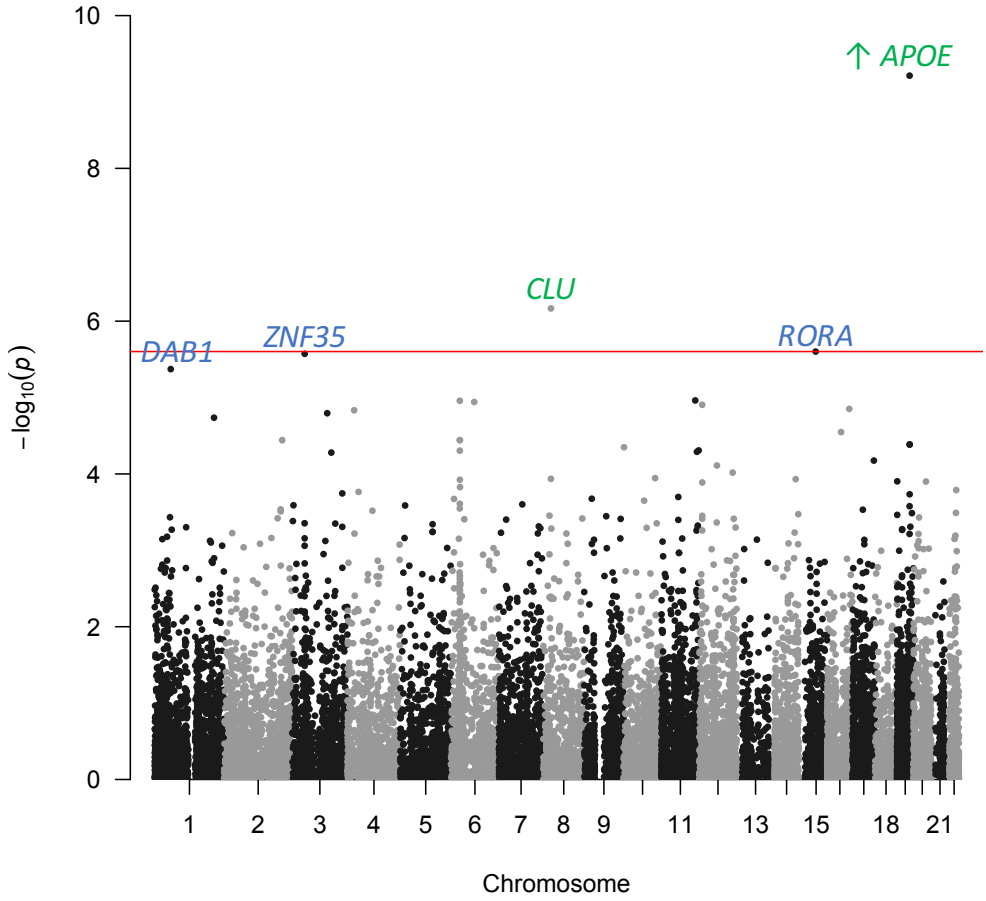


Figure 7.1: Manhattan Plot for the POLARIS Gene-Based Analysis in Imputed GERAD data

The gene-wide significant genes found when using the MAGMA-PCA approach in the GERAD imputed data are shown in Table 7.6 and the gene-wide significant genes using MAGMA-SUMMARY in the IGAP data (containing GERAD SNPs only) are seen in Table 7.7. The overlap of gene-wide significant genes between the three approaches; POLARIS, MAGMA-PCA and MAGMA-SUMMARY is seen in Figure 7.2. All genes found using MAGMA-PCA are also found by POLARIS and MAGMA-SUMMARY, and the MAGMA-SUMMARY approach finds 12 additional gene-wide significant genes.

Table 7.6: Gene-Wide Significant Genes from MAGMA-PCA Gene-based Analysis in AD Imputed Data

			MAGMA-PCA in GERAD
Chr	Gene	No. of SNPs	P-value
19	*BCL3*	5	2.7×10^{-8}
19	*PVRL2*	92	4.9×10^{-35}
19	*TOMM40*	20	1.6×10^{-48}

Table 7.7: Gene-Wide Significant Genes from MAGMA-SUMMARY Gene-based Analysis in IGAP Data (GERAD SNPs only)

			MAGMA-SUMMARY in IGAP
Chr	Gene	No. of SNPs	P-value
1	*CR1*	63	4.2×10^{-8}
2	*BIN1*	94	2.1×10^{-7}
8	*CLU*	13	1.0×10^{-11}
11	*SPI1*	16	1.8×10^{-6}
11	*MS4A2*	11	2.5×10^{-7}
11	*MS4A4E*	48	6.6×10^{-8}
11	*PICALM*	117	5.2×10^{-9}
19	*PVR*	16	1.4×10^{-8}
19	*CEACAM19*	13	2.0×10^{-6}
19	*BCL3*	4	2.4×10^{-20}
19	*PVRL2*	42	6.8×10^{-24}
19	*TOMM40*	6	7.3×10^{-32}
19	*APOC4*	8	3.8×10^{-9}
19	*APOC4-APOC2*	8	3.8×10^{-9}
19	*CLPTM1*	47	2.1×10^{-8}
19	*MARK4*	124	2.4×10^{-10}
19	*EXOC3L2*	19	9.7×10^{-7}

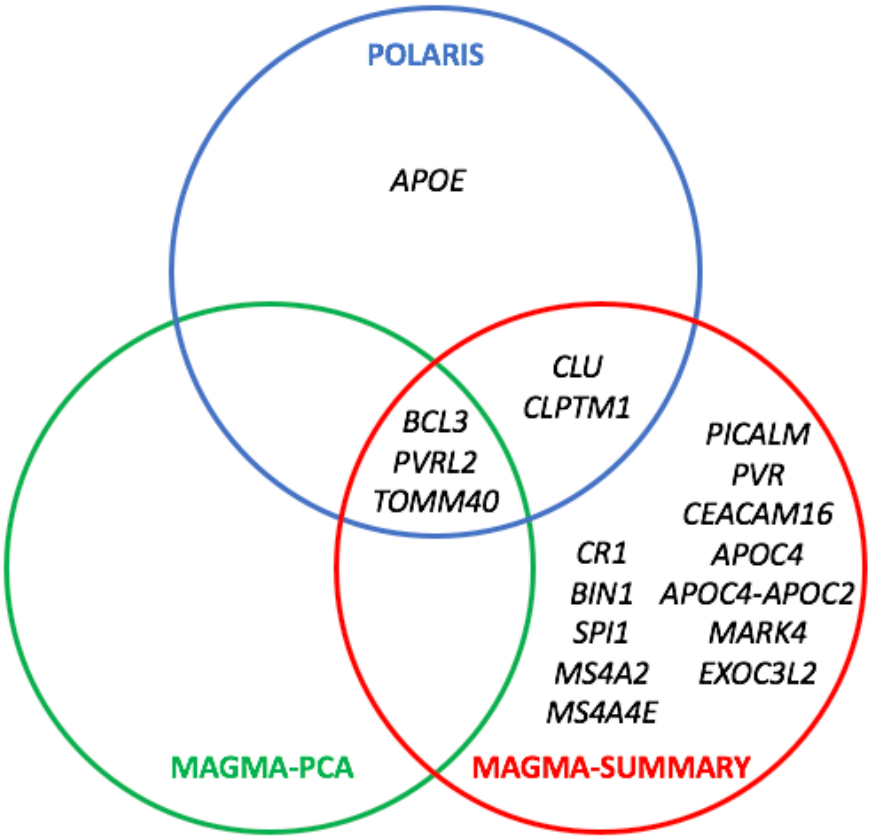


Figure 7.2: Venn Diagram Displaying the Overlap of Gene-Wide Significant Hits from POLARIS, MAGMA-PCA and MAGMA-SUMMARY

This analysis was repeated using a window around the gene of 35kb upstream and 10kb downstream since this was found to improve power, as transcriptional regulatory elements are likely to be contained within this window [76] (see Chapter 3). The 12 gene-wide significant genes from this analysis are seen in Table 7.8. Again, a large number of genes reside on chromosome 19, these are likely influenced by the large effect of *APOE*. Four novel genes have been identified from this analysis; *PPARGC1A*, *SCARA3*, *RORA* and *ZNF423* [107]. The *PPARGC1A* gene has been linked to energy metabolism and the generation of amyloid beta plaques [108] and has potential relevance to the human aging process. The *SCARA3* gene overlaps the gene *CLU* which has previously been identified as being associated with AD [19], and has been found to be associated with total brain volume [109], so is not actually a novel finding. The *RORA* gene has strong links with AD associated genes and genes which are differentially expressed in the hippocampus [110]. The *ZNF423* gene interacts with genes known to be associated with AD [111]. Therefore,

these novel genes seem to be sensible candidate genes for AD. Again, these gene-based results are plotted on a Manhattan plot in Figure 7.3.

Table 7.8: Gene-Wide Significant Genes from POLARIS Gene-based Analysis in AD Imputed Data Using a Gene Window (35kb upstream and 10kb downstream)

			POLARIS		
Chr	Gene	No. of SNPs	β	SE	P-value
4	*PPARGC1A*	480	0.877	0.1851	2.2×10^{-6}
8	*SCARA3*	240	0.526	0.1064	7.8×10^{-7}
15	*RORA*	1813	0.334	0.0674	7.4×10^{-7}
16	*ZNF423*	1056	0.551	0.1163	2.1×10^{-6}
19	*BCL3*	88	0.377	0.0674	4.2×10^{-9}
19	*CBLC*	50	0.605	0.1161	1.8×10^{-7}
19	*BCAM*	71	0.556	0.0543	1.4×10^{-24}
19	*PVRL2*	160	0.546	0.0299	9.4×10^{-75}
19	*TOMM40*	108	0.500	0.0298	3.4×10^{-63}
19	*APOE*	55	0.520	0.0315	4.4×10^{-61}
19	*APOC1*	34	0.475	0.0315	1.5×10^{-51}
19	*APOC4-APOC2*	62	0.615	0.0871	1.6×10^{-12}

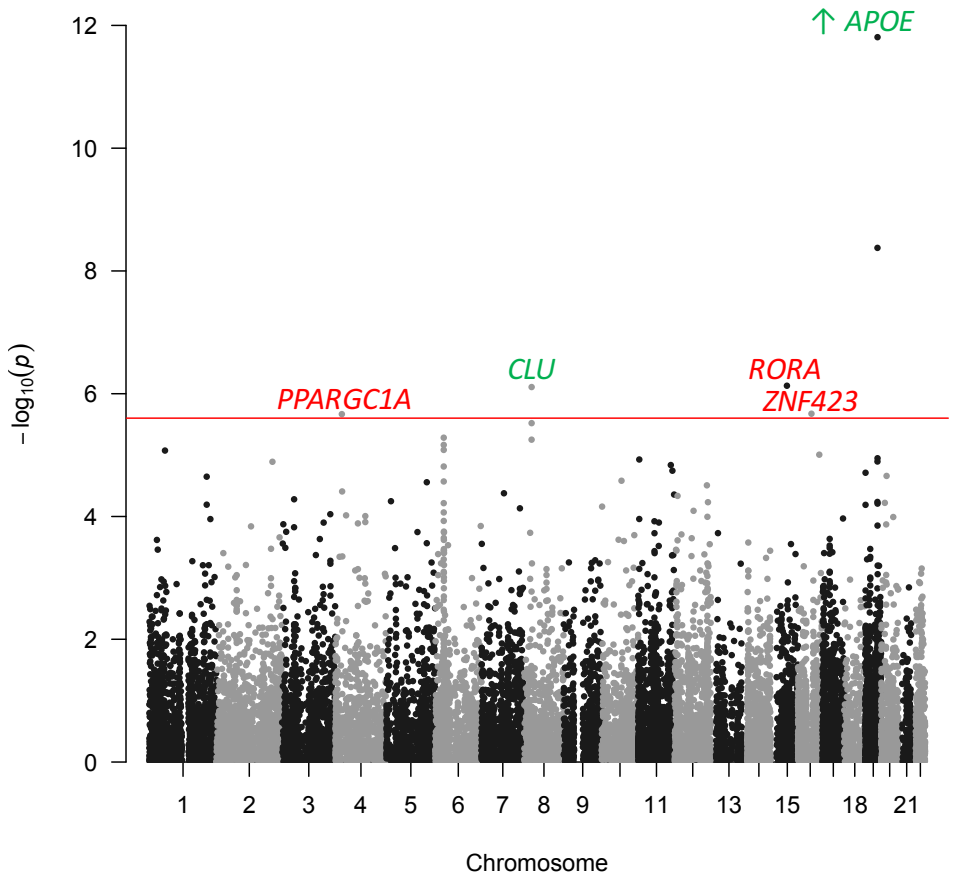


Figure 7.3: Manhattan Plot for the POLARIS Gene-Based Analysis in Imputed GERAD data Using a Gene Window 35kb Upstream and 10kb Downstream

7.3.1.1 Correlation Between P-values and the Number of SNPs in a Gene

The issue of bias in the estimation of set-based p-values caused by gene size is insufficiently tackled by most available methods [84][95]. Larger genes harbour a larger number of SNPs, and if each SNP has a small inflation in p-value due to, for example, unaccounted stratification, then these large genes will show greater accumulated inflation. To assess whether this is an issue in POLARIS, the phenotypes in GERAD data were permuted to create 1,000 simulations, and for each gene the empirical p-value (the proportion of gene-based p-values less than 0.05) was computed. The correlation between the number of SNPs per gene and the empirical p-value of each gene in AD data was then determined. No evidence (r=0.0009, p=0.9096) of a correlation between the number of SNPs in a gene and the gene p-value for the POLARIS method was found. Therefore, associations with

disease observed in larger genes is not simply due to a greater number of SNPs in the gene. Similarly, we observed no evidence (r=-0.00321, p=0.6977) of an inflation in p-value for increasing gene size using MAGMA-PCA on GERAD data. When considering the correlation between the IGAP gene-based p-value and set size, we observe a statistically significant negative correlation (r=-0.083, $p< 2.2 \times 10^{-16}$) when MAGMA-SUMMARY is used on summary data, indicating that the higher the number of SNPs, the lower the set-based p-value.

7.3.1.2 Conserved Regions

Loss of Function (LoF) genes from The Exome Aggregation Consortium (ExAC)

As discussed in Section 3.3.1.4, GWAS data was interrogated for genes that are evolutionary constrained, probably due to functional importance. It was assessed whether there was enrichment for either LoF intolerant genes using the POLARIS gene-based results in both genotyped and imputed data.

Table 7.9 show the contingency tables at different p-value thresholds, 0.05 and 2.5×10^{-6} respectively, for LoF genes based on the POLARIS gene-based results in genotype data. For the 14,620 genes from the analysis, there is no evidence that the genes found here are constrained for both the nominal and gene-wide p-value threshold (p=0.7786 and p=0.5665 respectively). The Fisher's exact test is used for the gene-wide p-value threshold, since the cell counts are small.

Table 7.9: Number of LoF Genes from the Exome Aggregation Consortium, Genotype Data

	$p \leq 0.05$	$p > 0.05$
in LoF	162	2595
out LoF	678	11185

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	1	2756
out LoF	3	11860

(ii) $p \leq 2.5 \times 10^{-6}$

The analysis presented in Table 7.9 does not adjust for potential correlation between genes. Since the same SNPs may be assigned to multiple genes, it is not possible to assume independence between genes. Therefore, the analysis is repeated removing genes

within 250kb of one another, retaining the most significantly associated gene. The results for the gene-based analysis are shown in Table 7.10. There are 563 genes which do not overlap, these show no enrichment for genes in conserved regions at either the nominal or gene-wide p-value threshold (p=0.9166 and p=1 respectively).

Table 7.10: Number of LoF Genes from the Exome Aggregation Consortium, Genotype Data, No Overlapping Genes

	$p \leq 0.05$	$p > 0.05$
in LoF	53	44
out LoF	249	217

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	0	97
out LoF	2	464

(ii) $p \leq 2.5 \times 10^{-6}$

Similarly, the results for the 17,072 genes from the POLARIS gene-based analysis in imputed data are seen in Table 7.11. There is no evidence that the genes were in constrained regions based on nominal and gene-wide significance p-values (p=0.07 and p=0.28).

Table 7.11: Number of LoF Genes from the Exome Aggregation Consortium, Imputed Data

	$p \leq 0.05$	$p > 0.05$
in LoF	380	2560
out LoF	1657	12475

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	2	2938
out LoF	4	14128

(ii) $p \leq 2.5 \times 10^{-6}$

Again, this analysis was repeated for non-overlapping genes from the gene-based analysis in the imputed data. The results are shown in Table 7.12, for the 557 non-overlapping genes, there is no evidence that the genes are enriched in conserved regions at either the nominal or gene-wide p-value thresholds (p=0.9052 and p=0.3505 respectively).

Table 7.12: Number of LoF Genes from the Exome Aggregation Consortium, Imputed Data, No Overlapping Genes

	$p \leq 0.05$	$p > 0.05$
in LoF	75	33
out LoF	317	132

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	1	107
out LoF	1	448

(ii) $p \leq 2.5 \times 10^{-6}$

These results indicate that there is no significant increase in the number of significant genes in conserved regions. This finding is opposite to that found in Schizophrenia [102], where authors observed that common GWAS signals are highly enriched among genes

under strong selection pressures.

Conserved Noncoding Sequences (CNS)

The contingency tables showing whether the genes from the POLARIS gene-based analysis reside in CNS are seen in Tables 7.13 for both nominal and gene-wide p-value thresholds. The cell counts are small for both tables and therefore a Fisher's exact test was used to determine whether there was an enrichment of genes in CNS. There was no evidence of enrichment for either the nominal or gene-wide p-value threshold (p=0.4026 and p=1, respectively).

Table 7.13: Number of Genes in Conserved Noncoding Sequences, Genotype Data

	$p \leq 0.05$	$p > 0.05$
in LoF	0	27
out LoF	840	13753

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	0	27
out LoF	4	14589

(ii) $p \leq 2.5 \times 10^{-6}$

The analysis considering only non-overlapping genes from the gene-based analysis in genotype data is shown in Table 7.14. For the 563 non-overlapping genes, there is no evidence of genes being enriched in CNS at either p-value threshold (p=1 for both).

Table 7.14: Number of Genes in Conserved Noncoding Sequences, Genotype Data, No Overlapping Genes

	$p \leq 0.05$	$p > 0.05$
in LoF	0	0
out LoF	302	261

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	0	0
out LoF	2	561

(ii) $p \leq 2.5 \times 10^{-6}$

The same analysis was carried out for the 17,072 genes from the POLARIS gene-based analysis in the imputed data, the results are seen in Table 7.15. Again, a Fisher's exact test is used for both tables due to the small cell counts. No enrichment of genes in CNS is observed for either the nominal or gene-wide p-value threshold (p=0.137 and p=1, respectively).

Table 7.15: Number of Genes in Conserved Noncoding Sequences, Imputed Data

	$p \leq 0.05$	$p > 0.05$
in LoF	6	22
out LoF	2031	15013

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	0	28
out LoF	6	17038

(ii) $p \leq 2.5 \times 10^{-6}$

Table 7.16 shows the analysis considering only non-overlapping genes. For these 557 genes, there is no evidence of an enrichment of genes in CNS at either the nominal or gene-wide p-value threshold (p=0.5586 and p=1 respectively).

Table 7.16: Number of Genes in Conserved Noncoding Sequences, Imputed Data, No Overlapping Genes

	$p \leq 0.05$	$p > 0.05$
in LoF	3	0
out LoF	389	165

(i) $p \leq 0.05$

	$p \leq 2.5 \times 10^{-6}$	$p > 2.5 \times 10^{-6}$
in LoF	0	3
out LoF	2	552

(ii) $p \leq 2.5 \times 10^{-6}$

Across all these different analyses, there is consistently no enrichment at the gene-wide p-value threshold for genes in either Loss of Function (LoF) regions or Conserved Noncoding Sequences (CNS). The analyses considering non-overlapping genes show consistent results, suggesting there is little correlation between genes, or that this correlation has little impact on the chi-squared test. It is anticipated that there will no enrichment of genes associated with AD in LoF or CNS since AD is a post-reproductive disorder.

7.3.2 POLARIS Pathway Analysis

The POLARIS pathway scores were computed for the eight pathways previously found to be associated with AD [23][24], see Table 7.17. Both self-contained and competitive tests are presented, where a self-contained test demonstrates whether the pathway is associated with AD, whereas the competitive test demonstrates whether a pathway shows an association with AD which is independent of the baseline level of association. The p-values from the original ALIGATOR analysis [61] are also presented in the table for comparison. A pathway is considered as significant if the p-value is below 3.125×10^{-3}, using a Bonferroni correction $(0.05 \div (8 \times 2))$ [71]. All eight pathways are significant in the

self-contained analysis, and four pathways withstand the adjustment for baseline association; these are the immune response, cholesterol transport, hematopoietic cell lineage and the clathrin pathways. The additional pathways found to be associated with AD demonstrate the power of the POLARIS method from utilising all available data. The number of significant pathways from the competitive test is greater using POLARIS compared to using PRS or MAGMA. ALIGATOR p-values are consistently lower than those from POLARIS, this is because ALIGATOR selects genes with at least one significant SNP whereas pathways calculated using POLARIS include all SNPs with a p-value less than 0.5 so is more influenced by noise.

Table 7.17: AD Associated Pathways Calculated Using POLARIS in GERAD data

Pathway	Beta	SE	P_{sc}	P_c	ALIGATOR p-value
1. Immune response	0.2754	0.0305	1.57×10^{-19}	4.30×10^{-10}	0.00266
2. Regulation of endocytosis	0.1295	0.0301	1.73×10^{-5}	0.0279	0.0002
3. Cholesterol transport	0.1428	0.0305	2.73×10^{-6}	8.55×10^{-5}	0.00024
4. Hematopoietic cell lineage	0.1560	0.0299	1.76×10^{-7}	6.27×10^{-6}	0.00007
5. Proteasome-ubiquitin activity	0.1070	0.0305	0.0004	0.0285	0.00929
6. Reactome hemostasis	0.1317	0.0301	1.18×10^{-5}	0.0434	0.00785
7. Clathrin	0.1734	0.0306	1.42×10^{-8}	0.0002	0.00038
8. Protein folding	0.0949	0.0301	0.0016	0.0471	0.00634

The correlations between all eight POLARIS pathway risk scores is shown in Table 7.18. Again, the pathway numbers correspond to those in Table 7.17 and the pathway risk scores are adjusted for population stratification by regressing the scores against principal components, and testing the pairwise correlation between the residuals from this model. Similarly to before, all pathways are strongly correlated with one another, although the actual correlation coefficients are very close to zero. The small p-values are due to the large number of individuals (N=13,164). The immune response pathway is most strongly correlated with all other pathways. The immune response and reactome hemostasis pathways have the largest positive correlation coefficient (r=0.3214).

Table 7.18: Correlations Between AD Associated Pathways, where Correlations were Calculated Using Individual PRS Estimated With POLARIS, where POLARIS is adjusted for population stratification *(Pathway Numbers Correspond to those in Table 7.17)*

Corr. Coeff (p-value)	1	2	3	4	5	6	7	8
1	1 (0)	0.1659 ($< 2.2 \times 10^{-16}$)	0.0861 ($< 2.2 \times 10^{-16}$)	0.1531 ($< 2.2 \times 10^{-16}$)	0.0815 ($< 2.2 \times 10^{-16}$)	0.3214 ($< 2.2 \times 10^{-16}$)	0.1326 ($< 2.2 \times 10^{-16}$)	0.0732 ($< 2.2 \times 10^{-16}$)
2		1 (0)	0.0524 (1.8×10^{-9})	0.0237 (0.0064)	0.0335 (0.0001)	0.1070 ($< 2.2 \times 10^{-16}$)	0.2137 ($< 2.2 \times 10^{-16}$)	0.0266 (0.0023)
3			1 (0)	0.0735 ($< 2.2 \times 10^{-16}$)	0.0464 (1.0×10^{-7})	0.0557 (1.6×10^{-10})	0.0948 ($< 2.2 \times 10^{-16}$)	0.0883 ($< 2.2 \times 10^{-16}$)
4				1 (0)	0.0059 (0.4962)	0.0791 ($< 2.2 \times 10^{-16}$)	0.0547 (3.3×10^{-10})	0.0389 (8.0×10^{-6})
5					1 (0)	0.0552 (2.3×10^{-10})	0.0419 (1.6×10^{-6})	0.0574 (4.3×10^{-11})
6						1 (0)	0.1441 ($< 2.2 \times 10^{-16}$)	0.0514 (3.6×10^{-9})
7							1 (0)	0.0451 (2.3×10^{-7})
8								1 (0)

7.4 Discussion

This chapter aimed to apply the POLARIS methodology introduced in Chapter 6 to the real AD data as both a gene and pathway based analysis.

Initially, a gene-based analysis was performed using the genotype GERAD data as the test set and IGAP data excluding the GERAD subjects was used to inform the analysis. POLARIS determined more statistically significant genes than the MAGMA-PCA method in the same GERAD data, since it uses the additional IGAP set. MAGMA-SUMMARY in the IGAP data has determined the largest number of gene-wide significant genes, although, as discussed in Section 3.3.3, this method seems to have unusually large power in summary statistics compared to the original genotype data. POLARIS does not identify any novel genes here, since this only considers genotype data, despite being informed with IGAP-noGERAD effect sizes. Perhaps the lack of novel genes identified could be due to heterogeneity between studies in IGAP, effects apparent in GERAD may be cancelled out by opposing effects in the other consortia.

The gene-based results in the HRC imputed GERAD data find a larger number of gene-wide significant genes compared to the analysis in the genotype data, suggesting the increased power of using the imputed data. No novel genes were determined until a

gene window was expanded around the gene, 35kb upstream and 10kb downstream. This reiterates the finding that the use of a window around the gene includes more informative SNPs and hence enhances the statistical power, since the window includes transcriptional regulatory elements in the gene. The three novel genes found to be associated with AD are *PPARGC1A*, *RORA* and *ZNF423*, all of which have credible biological relevance to AD [107].

PPARGC1A (peroxisome proliferator-activated receptor gamma co-activator 1alpha) is a transcriptional coactivator involved in a wide range of cellular and physiological functions. It is part of the PGC-1 family of transcriptional coactivators that mainly regulate mitochondrial biogenesis to in turn regulate the cellular energy metabolism [112]. The gene product, PGC-1, is an interacting partner of a wide range of nuclear receptors and transcription factors. It is associated with a wide-range of biological processes (Gene Ontology, `http://www.ensembl.org`), including response to a variety of cellular and external stimuli, cellular glucose homeostasis, circadian rhythm, regulation of neuron apoptosis, etc. Previous animal model work has shown that overexpression of hPGC-1 in APP23 mice improved spatial and recognition memory, along with significant reduction of Aβ deposition [108]. Furthermore, hPGC-1 overexpression also reduced the levels of proinflammatory cytokines and microglial activation [113][108]. This suggests a direct link with recent genetic evidence of microglia-mediated innate immune response involvement in AD [22]. In addition, an activation of PGC-1 by EKR and p38 inhibitors have been shown to improve spatial and learning memory in Aβ-injected rats [114]. *PPARGC1A* has also been implicated in the pathogenesis of other neurodegenerative disorders, namely Huntington's and Parkinson's diseases [115].

Retinoic acid receptor-related orphan receptor alpha (*RORA*) is a nuclear hormone receptor and is involved in a variety of functions; such as circadian rhythm, cholesterol metabolism and inflammation [116]. *RORA* binds to genomic regions of transcription start sites of more than 3,000 genes in human monocytic and endothelial cell lines [117]. *RORA* and *PPARGC1A* are close biological partners, with PGC-1 stimulating the expression of a number of clock genes through the coactivation of the ROR family of orphan nuclear receptors [118]. *RORA* has been shown to be linked to other genes previously implicated

in AD [111] and has also been implicated in a large number of neuropsychiatric disorders, such as post-traumatic stress disorder [119]. Furthermore, *RORA* trans-activates Il-6 and is thought to be neuro-protective in astrocytes and anti-inflammatory in peripheral tissues [120]. The two genes, *RORA* and *PPARGC1A* provide further evidence of the involvement of inflammation in the pathogenesis of AD.

Finally, *ZNF423* is a nuclear protein that belongs to the Kruppel-like C2H2 zinc finger proteins. *ZNF423* directs bone morphogenetic protein (BMP)-dependent signalling activity and aberrant forms impede B cell differentiation [121]. Furthermore, an increased gene-expression of *ZNF423* has been associated in patients with systemic lupus erythematosus pointing to an impaired function of B cells in human mesenchymal stem cells [122]. *ZNF423* resides in an AD-specific protein network [111]. *ZNF423* is likely involved in DNA damage repair [123]. Previously, it also has been shown that missense and LoF variants are likely pathogenic for abnormality of brain morphology, Joubert syndrome and Nephronophthisis with autosomal dominant or autosomal recessive inheritance (`www.omim.org,https://www.ncbi.nlm.nih.gov/clinvar/`). These disorders present with a range of phenotypic characteristics, with the central nervous system being affected too (more specifically the cerebellar vermis). In nur12 mouse model (with introduced nonsense mutation in exon 4 of the mouse Zfp423 gene), [124] observed loss of the corpus callosum, reduction of hippocampus, and a malformation of the cerebellum reminiscent of patients with Dandy-Walker syndrome. Within the cerebellum, Zfp423 was observed to be expressed in both ventricular and external germinal zones. Loss of Zfp423 was also observed to lead to diminished proliferation by granule cell precursors in the external germinal layer and abnormal differentiation and migration of ventricular zone-derived neurons and Bergmann glia [124].

It was demonstated that POLARIS results are not inflated by set size (number of SNPs in the gene), this effect was also not observed using MAGMA-PCA but was seen using MAGMA-SUMMARY.

As has been observed previously, genes determined from the gene-based analysis show no enrichment in conserved regions, in either evolutionary constrained regions or in CNS.

This is expected in AD because it is a post-reproductive disorder.

POLARIS pathway risk scores were produced for the eight pathways previously found to be associated with AD [23][24]. All eight of the POLARIS pathway risk scores were found to be associated with AD from the self-contained test. Four of these pathways remain associated after adjustment for baseline association; these are the immune response, cholesterol transport, hematopoietic cell lineage and clathrin pathways. The strongest correlation is between the immune response and reactome hemostasis pathways.

A Python code has been made available at `github.com/BakerEA/POLARIS`, a copy of the code is in the Supplementary Material Section 11.3. This code is able to handle set-based analyses when the sets contain less than 200 SNPs. The code is command line based, includes the SNP to gene annotation, allele flipping, POLARIS and logistic regression.

POLARIS could also be used in a similar way to the regular PRS approach whereby all SNPs across the genome are included. This is computationally demanding, and is not available in the current software. A newer software is currently under development which is written in C and aims to accommodate a larger number of SNPs and speed up calculations.

POLARIS has the advantages that 1) it produces a risk score per person per set. This can help to identify subjects for clinical trials and further functional studies. 2) POLARIS successfully adjusts for LD between SNPs and therefore LD pruning is not required prior to analysis. 3) POLARIS is able to utilise additional data to improve power whilst maintaining a test of association in the original data. 4) POLARIS is not influenced by the number of SNPs in the set, large sets are not biased and more likely to be significant.

8 POLARIS: Polygenic LD-Adjusted Risk Score Whole Genome Based Approach

8.1 Introduction

PRS is most widely used to determine the overall polygenicity of disease by combining the effects of common SNPs which show some association to disease, but do not reach the genome-wide significance threshold ($p < 5 \times 10^{-8}$). PRS has been used to show the polygenic contribution in a number of different complex disorders, such as AD [38], Parkinson's disease [125] and SZ [37]. PRS unlocks the latent information and is therefore able to explain a larger amount of heritability compared to genome-wide significant SNPs only [37]. Studies have shown that including loci which do not reach genome-wide significance into the polygenic score have increased estimated heritability in AD [126]. Another motivation for using whole genome PRS is to predict disease case/control status or the trait liability of all SNPs, and it is expected that using all SNPs will have better discriminatory power than prediction based on set-specific risk scores.

POLARIS [96] has been shown to be a powerful set-based method compared to MAGMA approaches. POLARIS produces a risk score per individual per set of SNPs whilst adjusting for LD between SNPs. Like standard PRS [37], POLARIS is applicable to any set of SNPs, including all SNPs across the whole genome. In this Chapter the use of POLARIS as a whole genome approach is examined in the genotyped GERAD data, using IGAP-noGERAD (IGAP data excluding GERAD subjects) as weights in the score.

POLARIS uses the square inverse of the correlation matrix between SNPs in order to cor-

rect genotypes for LD, giving LD adjusted dosages. Clearly, using all SNPs in the genome will require the inversion of a large matrix (419,048 x 419,048 for genotyped GERAD data and 3,169,840 x 3,169,840 for imputed GERAD data) which is highly computationally demanding. The simplest approach to reduce the computational burden is to split all SNPs across the genome into chromosomes; it is assumed there is no LD between SNPs on different chromosomes (Chr1: 32,285 x 32,285 in genotyped GERAD data and 240,991 x 240,991 in the imputed GERAD data; Chr22: 6,301 x 6,301 in genotyped GERAD data and 42,381 x 42,381 in imputed GERAD data). The whole genome risk score is then simply a sum of all chromosomal risk scores. However, with the increasingly widespread use of imputed data, the number of SNPs per chromosome remains large and therefore this still has computational issues. Another approach to resolve the computational burden would be to use a sliding window of SNPs, although the windows must overlap and therefore, it is complicated to adjust for LD when there are multiple measures for each SNP. Due to the large dimensionality of the imputed GERAD data and to make results comparable with [21] and [38], only the genotyped GERAD data is used in this chapter. The approach which computes the POLARIS score per chromosome, and then sums the chromosomal scores is used; this is the simplest method since it does not require defining the LD region size.

LDpred [127] is an alternative method to POLARIS which also adjusts effect sizes for the LD structure of SNPs. LDpred uses a Bayesian approach which estimates the posterior mean effect sizes from the discovery data and the LD information in the SNPs. For large sample sizes, the β adjustment in LDpred corresponds to multiplying the effect sizes (βs) by the inverse of the correlation matrix. LDpred has two different models, the infinitesimal model which assumes that all SNPs are causal and the noninfinitesimal model which assumes some SNPs are not causal. The noninfinitesimal models use a p-value threshold for the inclusion of SNPs, here, the p-value threshold used is 1, so that all SNPs are included. The p-value thresholding is equivalent to that used in both POLARIS and PRS.

POLARIS [96] uses spectral decomposition of the SNP correlation matrix to adjust the individuals' allele counts for LD structure, however, this is equivalent to adjusting the

effect sizes (β) or weights which are used in the risk score. For this reason, POLARIS used on all SNPs across the whole genome is compared to LDpred.

In this chapter, POLARIS is evaluated by comparing results calculated using POLARIS with those found using unadjusted PRS [37] and LDpred [127] on simulated and real AD data. The effect size adjustment in POLARIS and LDpred are also compared.

PRS across the whole genome is often used for disease prediction; to determine whether the polygenic component of disease is able to predict whether a subject is a case or control. This has been investigated using standard PRS in pruned GERAD data (using the --clump option in PLINK, which prioritises SNPs which are most associated with disease, hereafter referred to as intelligent pruning), where the maximum prediction accuracy for AD is found by incorporating PRS using a p-value threshold of 0.5 into a model including the number of *APOE*-ϵ4 alleles, the number of *APOE*-ϵ2 alleles, a PRS including index SNPs from the 20 GWAS hits, age and sex, giving an Area Under the Curve (AUC) of 0.782 [38]. This has been calculated as 90% of the total prediction accuracy available, based on genetics alone [128]. The impact of using the POLARIS score in these prediction models is assessed to determine if POLARIS is able to improve upon prediction accuracy using just *APOE* effects and GWAS hits.

8.1.1 Objectives

This Chapter aims to:

- Compare the power of POLARIS to LDpred and unadjusted PRS in both simulated and real AD GERAD data, using effect sizes from IGAP-noGERAD .
- Compare the LD adjustment of the effect sizes (βs) in POLARIS and LDpred in simulated data. This is to observe the changes in βs after LD adjustment, in data where the LD structure is defined and therefore simple to interpret.
- Consider POLARIS as a whole genome PRS approach in real AD data.
- Discuss the computational issues with running POLARIS in large sets of SNPs and

consider possible solutions. POLARIS involves the inversion of the correlation matrix, which can be computationally expensive when the matrix is large.

- Assess whether POLARIS across the whole genome is able to predict AD case-control status. This prediction accuracy is compared to that using PRS in intelligently pruned data [38].

8.2 Materials and Methods

8.2.1 POLARIS

POLARIS [96] is a method which adjusts genotypes for LD, creating a dataset containing independent LD adjusted genotypes which can then be used to compute a polygenic score. POLARIS has been introduced as a set-based method, but is applicable to any set of SNPs, including all SNPs across the genome. The simplest approach to do this is to compute the POLARIS score for all SNPs on each chromosome, and then sum across the chromosomal scores to determine a whole genome risk score. This simple chromosomal approach was used in this chapter, since it does not require the definition of the size of the LD region. Due to the higher dimensionality of the imputed data, only the genotyped GERAD data is used in this analysis.

8.2.2 LDpred

LDpred [127] is a Bayesian approach which also adjusts the standard PRS for LD between SNPs. It does this by determining the posterior mean effect size for each SNP using a prior on effect sizes (βs) and estimating LD from an external reference panel.

When sample sizes are large, the LDpred β adjustment is simply done by multiplying the βs by the inverse of the correlation matrix, see Equation 8.1.

$$\tilde{\beta} \approx D^{-1}\beta \tag{8.1}$$

where $\tilde{\beta}$ are the LD adjusted effect sizes, D is the correlation matrix between SNPs estimated from the test set and β are the original effect sizes from the discovery set.

POLARIS produces risk scores whilst adjusting for LD between SNPs. The adjustment in POLARIS using the square inverse of the correlation matrix was compared to using an adjustment of the inverse of the correlation matrix; it was shown that the type I error was consistent, but POLARIS gave higher power, see Chapter 6. However, LDpred is expected to differ slightly from the results adjusting for LD using the inverse of the correlation matrix, since if only a proportion of SNPs are assumed to be causal, a Markov Chain Monte Carlo (MCMC) method is used, as an explicit solution is too complicated [127].

LDpred avoids the computational issue of handling large matrices by performing the LD adjustment in smaller regions, since the LD matrix is approximately block diagonal. Although, the user has to define the size of the region, this can be difficult and some LD between SNPs could be missed if the defined region is too small.

LDpred takes the effect sizes (βs) from the discovery data, and uses the genotype data available in the test set in order to estimate the LD structure between SNPs.

8.2.3 Comparison between POLARIS, PRS and LDpred

To understand detailed differences and similarities between POLARIS, PRS and LDpred, these approaches were used on extreme simulated data and real data LD patterns between SNPs. Both type I and type II errors were investigated by simulating null effects and introducing some association to the SNPs.

Genotype (test) data and summary statistic (discovery) data were simulated in order to compare the power between POLARIS and LDpred. A dataset was simulated which was then randomly split into discovery and test sets. The summary statistics for each SNP in the discovery set were computed. The following LD structure was simulated: 4 LD Blocks of 10 SNPs each, and 60 independent unassociated SNPs. Block 1 has pairwise $r^2 = 0.2$, Block 2 has pairwise $r^2 = 0.4$, Block 3 has pairwise $r^2 = 0.6$, and Block 4 has pairwise

$r^2 = 0.8$, all 40 SNPs in LD with OR $\sim$ N(1.02, 0.36) (OR from a Normal Distribution with mean 1.02 and variance 0.36). The mean and variance for the sampled effect sizes were calculated from all SNPs in the IGAP data [20]. The LD structure for this simulated data is seen in Figure 8.1.

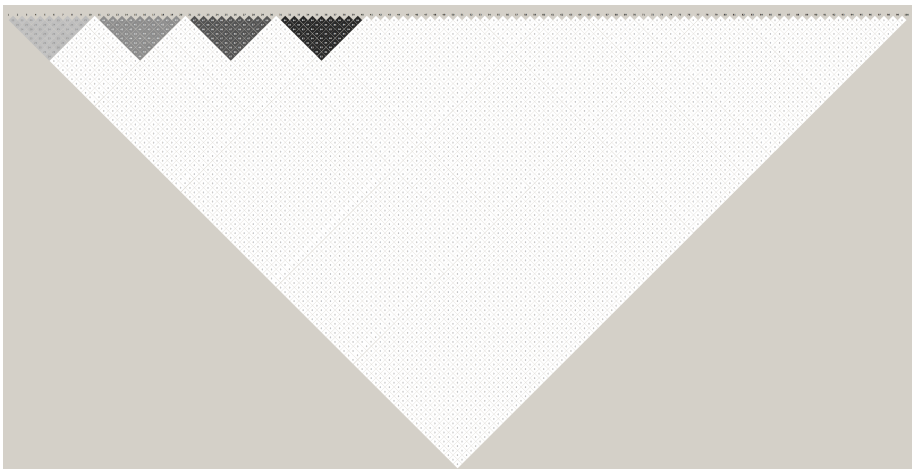


Figure 8.1: LD Plot for 100 SNPs in Complex LD Simulations

Only POLARIS and LDpred were compared in this simulated scenario; PRS requires LD pruning and since only SNPs in LD are associated with disease, these would be removed by LD pruning. So PRS would show very low power compared to the other methods.

The sample size of both the test and discovery dataset was varied in order to determine the influence of sample size on all methods. Test set sample sizes were N=1,000 and N=10,000 and discovery set sample sizes were N=10,000 and N=50,000. In both the discovery and test datasets, 30% of the sample size were cases.

A total of 1,000 simulations were performed. The power to detect the association between the score and disease is calculated as the proportion of p-values from the 1,000 simulations which were below a given p-value threshold; the p-value thresholds used were p=0.05, 0.01 and 0.001.

To ensure LDpred [127] was comparable to POLARIS, the LD structure was estimated from the test dataset to adjust effect sizes in the discovery set, using an LD radius of 100 SNPs (the number of SNPs on each side of the focal SNP for which LD should be adjusted) and other parameters as default. Both the infinitesimal model and p-value threshold of 1

are presented for simulated data, where the infinitesimal model assumes that all markers are causal.

The β adjustment between POLARIS and LDpred were plotted in order to compare the adjustment between the two approaches. The first simulation was taken for the case where the test set N=1,000 and the discovery set N=10,000 and where the test and discovery sample sizes are both 10,000. Each of the 100 SNPs were plotted along the x-axis and the original and adjusted βs were plotted on the y-axis. Vertical lines were used to demonstrate the differences between the strength of LD in each of the LD blocks, so the degree of adjustment is comparable. The LDpred approach substantially reduces the magnitude of the β values, so for these plots, the adjusted βs have been scaled by the gradient between the original and adjusted βs in the independent SNPs.

POLARIS, PRS and LDpred were additionally compared on all SNPs in both pruned and unpruned real AD data. The GERAD [19] data (GWAS) data (3,332 cases, 9,832 controls) were used as the test dataset and the IGAP [20] data (17,008 cases, 37,154 controls) excluding GERAD subjects were used as the discovery data in order to inform all analyses with association effect sizes (βs). Data is pruned using an r^2 threshold of 0.2, and both random and intelligent pruning (--indep and -- clump options in PLINK [50][51] respectively) were considered, where intelligent pruning retains the most associated SNPs. The unpruned genotyped GERAD data contains 419,048 SNPs, the intelligently pruned data contains 60,510 SNPs and the randomly pruned data has 94,277 SNPs. All analyses excluded chromosome 19 to remove the large effect of *APOE*.

For this study, only directly genotyped SNPs from the GERAD data were used, since this makes the results directly comparable with [21] and [38]. The GERAD data and IGAP data excluding GERAD subjects have 419,048 SNPs in common.

It was necessary to ensure that SNP alleles were coded in the same direction across both the discovery (IGAP excluding GERAD subjects) and test (GERAD) datasets. This is due to the effect sizes for a SNP being relative to a specific allele, so this reference allele must be consistent across the two datasets. If alleles in the discovery set were coded in opposite direction to those in GERAD, the summary effect size for the SNP was inverted.

SNPs with alleles AT, TA, CG or GC were excluded since the direction of the effect could not always be determined when combining two studies. Of the SNPs in the discovery set, 103,356 matched those in GERAD, the remaining had effect sizes inverted and no SNPs were excluded due to ambiguity. An MAF filter of 0.01 was applied to the data.

The missing genotypes in real data were imputed as in PLINK [50][51], where missing genotypes are substituted by $2 \times MAF$ for each SNP. In the GERAD data, 0.0514% of genotypes required imputation.

The whole genome risk scores were computed using a number of different individual p-value thresholds for the inclusion of SNPs (p=0.0001, 0.001, 0.01 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1) and a logistic regression model was used to determine the overall association of the risk score with AD, adjusting for population covariates such as age, sex, ethnicity and PCs.

8.2.4 Prediction Modelling Using POLARIS

The ability of the whole genome POLARIS score to predict whether an individual has AD or not was investigated. The sensitivity, specificity and AUC are computed for the POLARIS scores with each individual p-value threshold. Sensitivity is the probability of correctly predicting a person has AD and the specificity is the probability of correctly predicting a person does not have AD. The AUC is found from a Receiver Operating Characteristic (ROC) curve which plots the sensitivity on the y-axis and 1-specificity on the x-axis; an AUC of 1 shows perfect prediction ability and a value of 0.5 represents no prediction accuracy, or the equivalent of randomly guessing whether a person will have AD or not. These values are computed using the pROC package in R software.

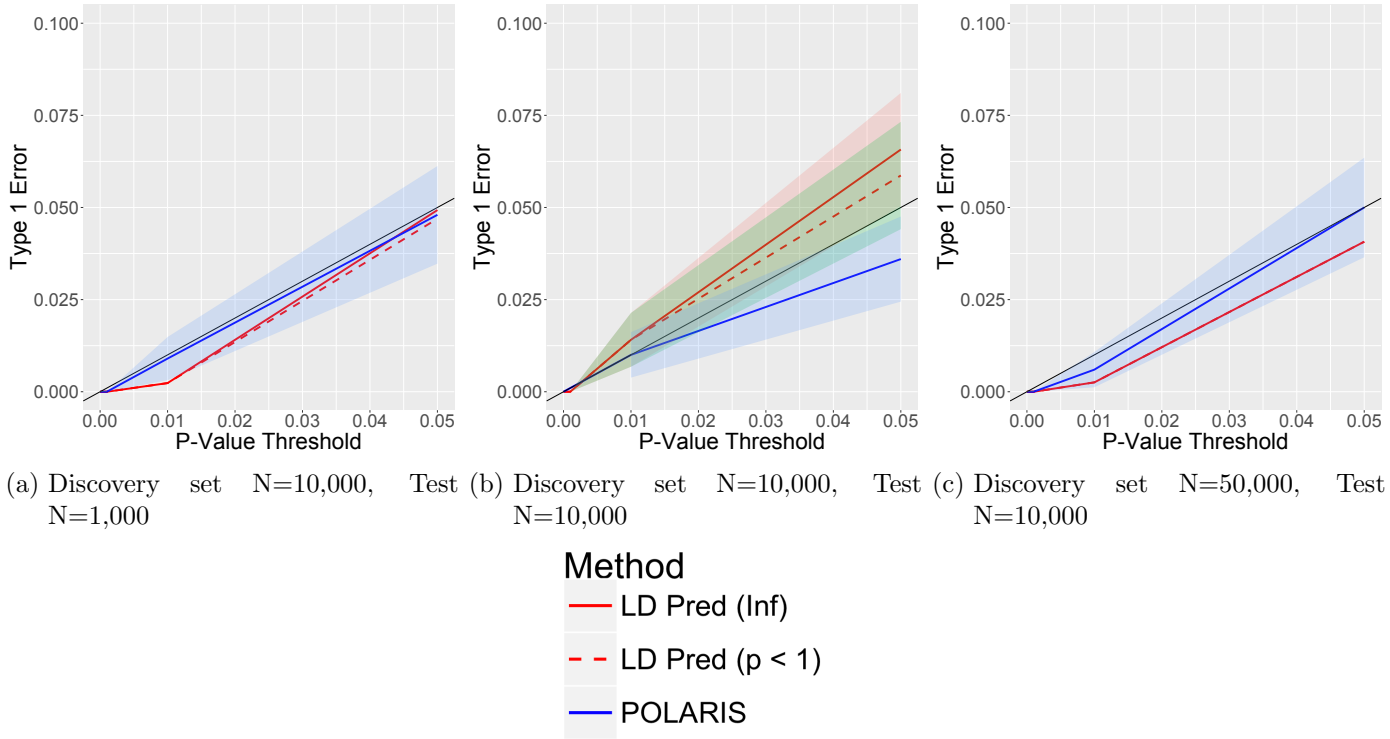
Prediction models included the POLARIS score along with a number of other covariates; the number of *APOE* $\epsilon 4$ alleles, the number of *APOE* $\epsilon 2$ alleles, the PRS score containing the 20 GWAS index SNPs (see Supplementary Table 11.2 for details), age and sex in order to compare with the PRS prediction analysis in intelligently pruned GERAD data [38].

8.3 Results

8.3.1 Comparison Between POLARIS, PRS and LDpred

8.3.1.1 Simulation Results

Firstly, the comparison between LDpred and POLARIS in simulated data is considered. The simulated data have 4 LD Blocks of 10 SNPs each, and 60 independent unassociated SNPs. Figure 8.2 displays the type I error for the different methods with the corresponding 95% CIs. In all type I error and power graphs, POLARIS is displayed as the blue line, the infinitesimal LDpred model is shown by the solid red line and the LDpred model using a p-value threshold of 1 is displayed by the dashed red line. Type I error is slightly higher in LDpred compared to POLARIS when the discovery set N=10,000, but the nominal value still resides within the 95% CI. The results for all sample sizes seem reasonable, with the LDpred type I error being slightly inflated when the test and discovery set have N=10,000.



(a) Discovery set N=10,000, Test N=1,000

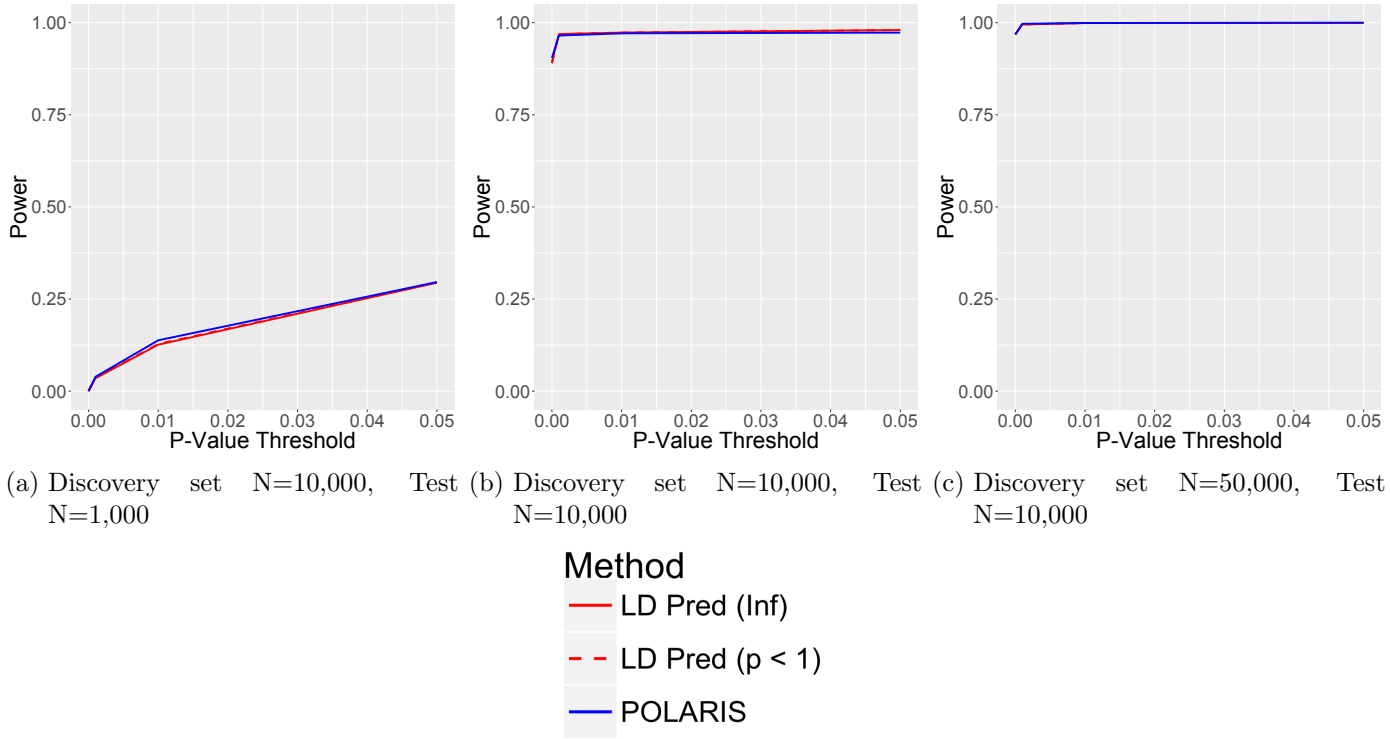
(b) Discovery set N=10,000, Test N=10,000

(c) Discovery set N=50,000, Test N=10,000

Figure 8.2: Type I Error Comparison Between POLARIS and LDpred: Simulation of 10 SNPs in LD with $r^2 = 0.2$, 10 SNPs in LD with $r^2 = 0.4$, 10 SNPs in LD with $r^2 = 0.6$, 10 SNPs in LD with $r^2 = 0.8$ and 60 independent, unassociated SNPs.

The power comparison between POLARIS and LDpred is shown in Figure 8.3. The two methods have very similar power for all simulated sample sizes.

It is evident that POLARIS and LDpred have almost identical power in all cases, and it is difficult to differentiate between the two methods. Figure 8.4 shows the same plots, but zoomed into a specific point on the y-axis in order to observe slight differences between POLARIS and LDpred. The power is slightly higher for POLARIS compared to LDpred, except when the test and discovery set N=10,000, this is the same case where LDpred has inflated type I error compared to POLARIS, so perhaps this explains the difference here.

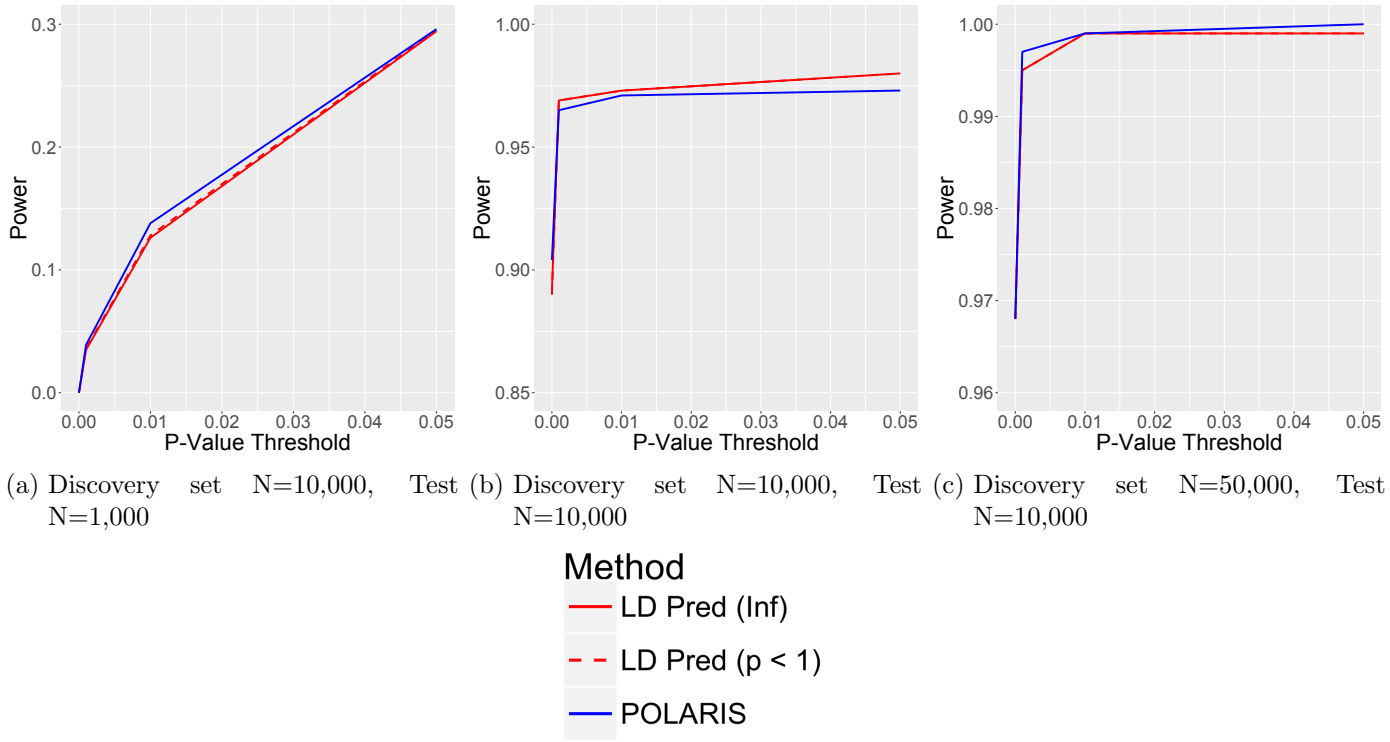


(a) Discovery set N=10,000, Test N=1,000

(b) Discovery set N=10,000, Test N=10,000

(c) Discovery set N=50,000, Test N=10,000

Figure 8.3: Power Comparison Between POLARIS and LDpred: Simulation of 10 SNPs in LD with $r^2 = 0.2$, 10 SNPs in LD with $r^2 = 0.4$, 10 SNPs in LD with $r^2 = 0.6$, 10 SNPs in LD with $r^2 = 0.8$, all 40 SNPs in LD Blocks have OR$\sim N(1.02, 0.36)$, and 60 independent, unassociated SNPs.



(a) Discovery set N=10,000, Test N=1,000

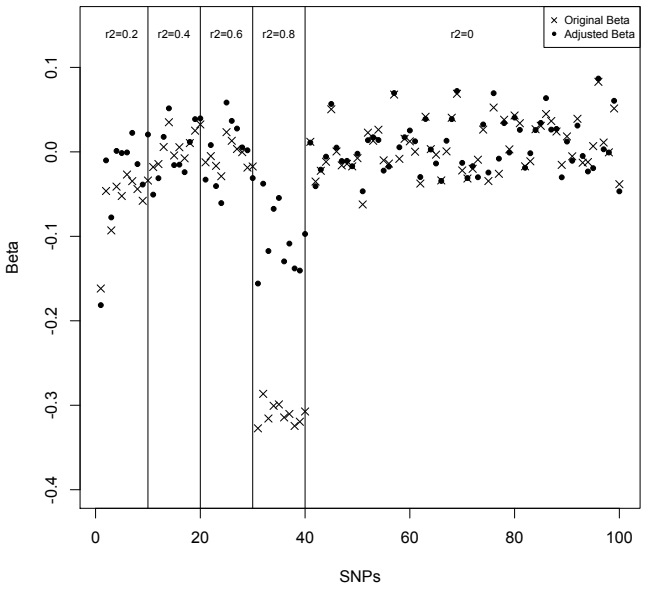
(b) Discovery set N=10,000, Test N=10,000

(c) Discovery set N=50,000, Test N=10,000

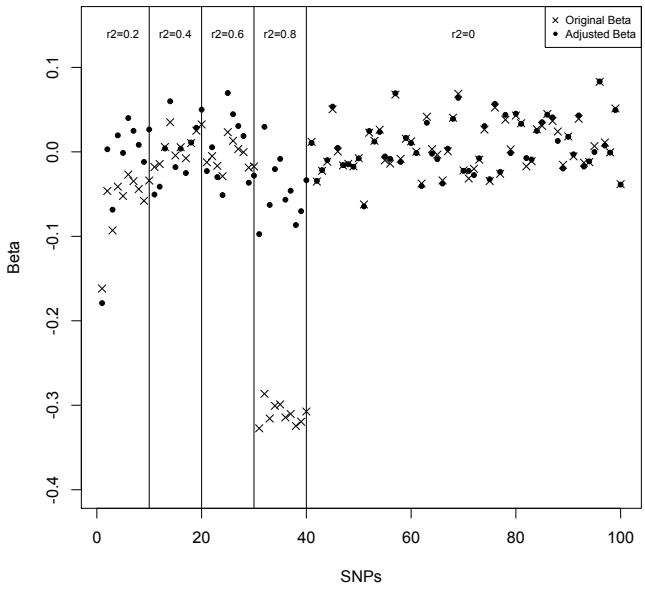
Figure 8.4: Power Comparison Between POLARIS and LDpred with a closeup of y-axis: Simulation of 10 SNPs in LD with $r^2 = 0.2$, 10 SNPs in LD with $r^2 = 0.4$, 10 SNPs in LD with $r^2 = 0.6$, 10 SNPs in LD with $r^2 = 0.8$, all 40 SNPs in LD Blocks have OR$\sim N(1.02, 0.36)$, and 60 independent, unassociated SNPs.

The β adjustment for LD in both POLARIS and LDpred for the first simulation from the 1,000 simulations is seen in Figure 8.5; the dots represent the original βs and the crosses are the LD adjusted βs. For LDpred, the noninfinitesimal model with $p < 1$ is used, although the results for the infinitesimal model are very similar (not presented here). LDpred reduces the magnitude of β, so the β adjustment plots have the adjusted βs scaled by the gradient between the original and adjusted β values for independent SNPs. The β adjustment in POLARIS and LDpred was investigated using a discovery set with N=10,000 and a test set with both N=1,000 and N=10,000 in order to estimate LD structure. For POLARIS, there is little to no adjustment in the 60 independent SNPs, with the amount of adjustment decreasing with increasing sample size of the test set. This is likely due to the LD between markers being better estimated in the larger set. For SNPs in LD, there is adjustment in the β values, with the degree of adjustment increasing for higher LD, again, this is more pronounced when the test set has a larger sample size. The 60 independent SNPs also have little to no adjustment using LDpred, which is expected. When the test

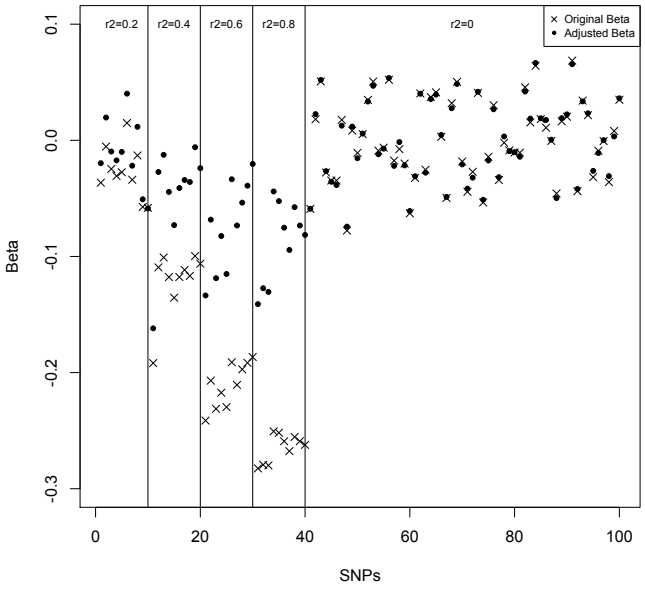
set is small (N=1,000), SNPs in LD have adjusted β values, with the largest adjustment in the LD block with the highest r^2 value, this is similar to the results for POLARIS. For a larger test set (N=10,000), the β adjustment in LDpred is far less pronounced, and almost absent for intermediate $r^2 = 0.4, 0.6$. This is counterintuitive; and the LDpred paper [127] does not suggest any reason why large sample sizes may affect the β adjustment, since large samples lead to the adjustment seen in Equation 8.1. The only explanation would be that the difference is caused by the numeric MCMC method used for the non-infinitesimal model.



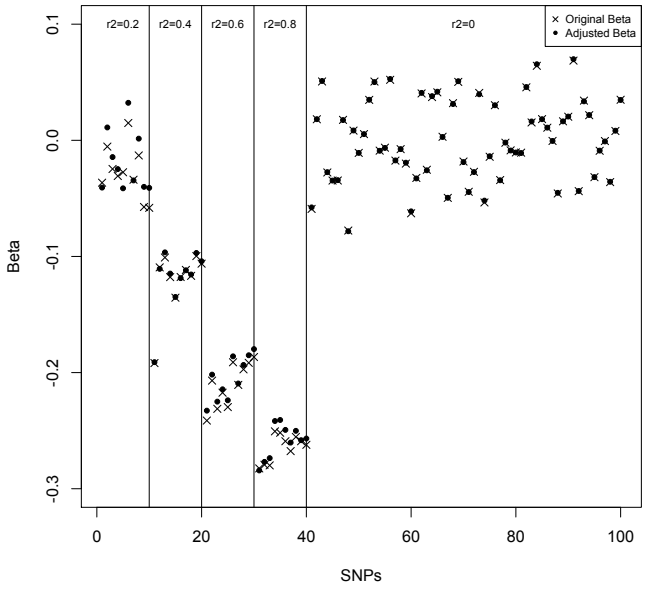
(a) **POLARIS:**
Discovery set N=10,000, Test N=1,000



(b) **LDPred:**
Discovery set N=10,000, Test N=1,000



(c) **POLARIS:**
Discovery set N=10,000, Test N=10,000



(d) **LDPred:**
Discovery N=10,000, Test N=10,000

Figure 8.5: Comparison Between β Adjustment in POLARIS and LDpred

8.3.1.2 Real AD Data Results

The overall p-values for the whole genome in the GERAD data using a logistic regression model of risk scores were compared between POLARIS (solid line), unadjusted PRS (dashed line), and LDpred (dotted line) depending on the individual p-value threshold used as criterion for the inclusion of SNPs. The comparison was performed for unpruned data (green lines in Figure 8.6), for intelligently pruned data, where the most associated SNPs are retained (r^2 threshold 0.2, red lines in Figure 8.6), and for randomly pruned data (r^2 threshold 0.2, blue lines in Figure 8.6). In order to assess the influence of the well-known strong association of *APOE*, the comparison excluded chromosome 19.

When data are either intelligently or randomly pruned for LD, the results for POLARIS, PRS and LDpred are very similar at all p-value thresholds. This is to be expected, as LD has been largely removed in these cases, so POLARIS and LDpred will make no or only small adjustments to the weights (βs). Intelligently pruned data show more significant results throughout compared to randomly pruned data, since highly associated SNPs are prioritised for inclusion. For larger inclusion thresholds, these data reach a stable maximum significance plateau at about p-value threshold p=0.5 in all methods; similar results for PRS were reported in [38]. When unpruned data are used, the results for PRS are more significant at all p-value thresholds than PRS in clumped and randomly pruned data on the whole genome. This is because PRS will be inflated by the LD between SNPs, by overaccounting for the effect of SNPs in LD. LDpred has increased overall significance in unpruned data, with higher significance levels than PRS in unpruned data. LDpred reaches a maximum overall significance at a p-value threshold of 0.1. For POLARIS, the overall significance increases with increasing p-value threshold, reaching a maximum beyond the values achieved for PRS and LDpred. This shows that POLARIS and LDpred unlock the information contained in SNPs which are not individually associated, but still have p-values below about 0.5, but POLARIS achieves a higher maximum overall significance. POLARIS, however, is sensitive to the signal-to-noise ratio; the green curve in Fig.8.6 decreases when SNPs with $p > 0.5$ are included. The reason for this may be that POLARIS increases the variance of the β distribution, and therefore, these unassociated SNPs have

a detrimental effect on the overall score compared to PRS. In an attempt to remove this effect, POLARIS has been altered and the whole genome analysis was rerun. The first approach was to set any βs which had a corresponding p-value above a particular threshold (i.e. p=0.8, 0.9) to zero. The second approach was to determine if the adjusted βs were larger than the original βs, and to return the adjusted value to the original if this was the case. Unfortunately, neither of these approaches removed the signal-to-noise sensitivity in POLARIS. The power for unpruned PRS is slightly higher than POLARIS when the p-value inclusion threshold is less than 0.1, this may be attributed to the inflated type I error in unpruned PRS compared to POLARIS. When lower p-value inclusion thresholds are used, the $-log_{10}$(p-values) are lower than when a higher inclusion threshold is used. This may reflect that there are fewer markers with small p-values, which only explain a limited fraction of disease heritability. PRS, POLARIS and LDpred are designed to incorporate the combined effect of SNPs which do not individually reach genome-wide significance in order to unlock this latent information and explain a larger proportion of the heritability [37].

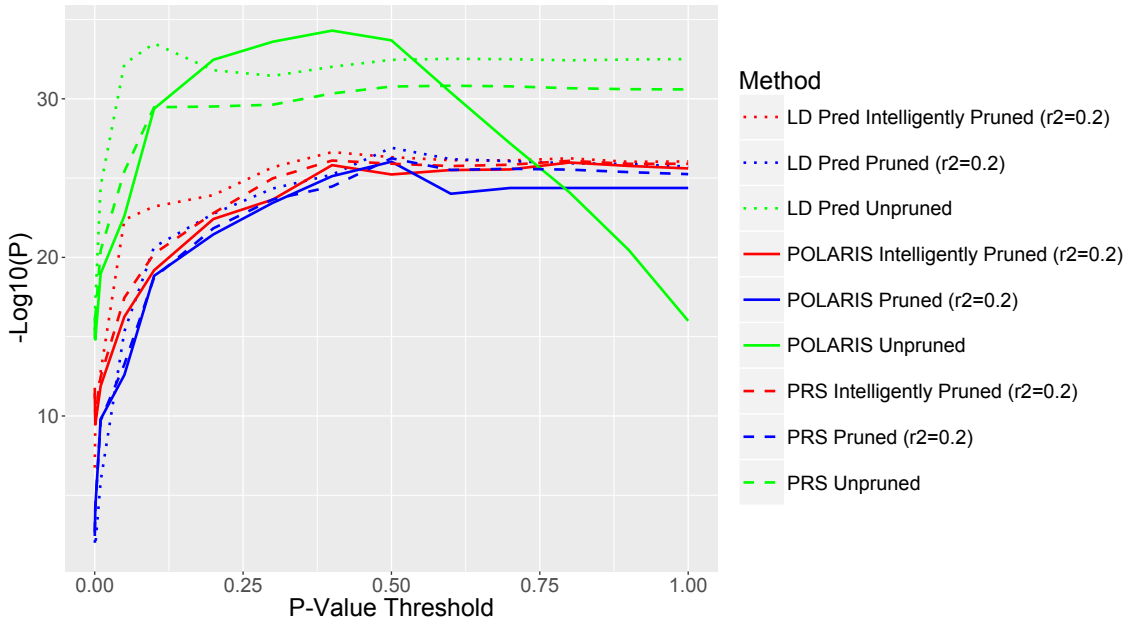


Figure 8.6: Plot of the $-log_{10}$(p-value) at different p-value inclusion thresholds for POLARIS, LDpred and PRS, in unpruned, intelligently pruned and clumped data.

8.3.2 Extensions to POLARIS

The original POLARIS methodology is extended such that it utilises the fact that the number of individuals, N, is often fewer than the number of SNPs, M, across the whole genome. An update to POLARIS decomposes the NxN matrix rather than the MxM matrix when $N < M$. The eigenvalues and eigenvectors of this NxN matrix are then adjusted to correspond with those for the MxM matrix. Eigenvectors which have eigenvalues < 0 are removed, thus computing the pseudoinverse of the correlation matrix. The results discussed previously do not use the pseudoinverse, perhaps this may reduce the signal-to-noise sensitivity observed in POLARIS.

POLARIS is able to use dosage data rather than genotype data. The expected dosage value is computed using Equation 8.2. This expected value is then used in the POLARIS score rather than the genotype.

$$E[genotype] = 2\times\text{P(dominant homozygote)}+1\times\text{P(heterozygote)}+0\times\text{P(recessive homozygote)} \tag{8.2}$$

8.3.2.1 POLARIS software

POLARIS has been implemented into a software written in Python, which can be found at `github.com/BakerEA/POLARIS`. This software is aimed at using POLARIS in a set-based framework, a maximum set size of 200 SNPs is able to be used in this software. The software reads data in the form of PLINK binary files and additionally annotates SNPs to genes, ensures effect sizes are with respect to the same alleles in the test and discovery sets, produces the POLARIS score and also finds the set association with disease.

Clearly, the currently available software is unable to undertake the whole genome analyses which have been carried out in this chapter. All analyses in this chapter were carried out using a Matlab code. This is a crude code which requires a substantial amount of data

preparation and Matlab is not freely available, and thus this code has not been made publically available. The extensions to POLARIS such as using dosage data rather than genotype data was implemented in an R code, but is not avaiable in the current software.

The current Python software has a maximum set size and therefore is not possible to be used for very large data or sets. Theoretically, as considered in this chapter, POLARIS is able to be used for any set of SNPs, ideally including all SNPs across the whole genome. The code used for the analyses in this chapter is not suitable for other researchers and therefore, a new code written in C is under development in order to allow POLARIS to be used across the whole genome. The issue with large sets is that a huge amount of Random Access Memory (RAM) is required to compute the square inverse of the correlation matrix and this is very slow as a large number of computations are required. The C code attempts to utilise the LAPACK [129] and ScaLAPACK [130] packages which optimise matrix inversion, and are able to run in parallel. This software is still undergoing development before it will be released for the use of other researchers.

8.3.3 Prediction Modelling Using POLARIS

The whole genome POLARIS score can be used to predict whether a subject has AD or not. The prediction ability of the POLARIS score to predict AD cases was assessed in the GERAD data.

The sensitivity is the probability of correctly predicting whether a person has AD and the specificity is the probability of correctly predicting that a person does not have AD.

The AUC is used to determine the prediction accuracy of the model. An AUC value of 1 is when the model is perfectly able to predict whether or not a subject has AD, and a value of 0.5 corresponds to randomly guessing whether or not a person has AD.

Tables 8.1-8.4 show the results for these prediction models. The four different tables correspond to a POLARIS score which includes different SNPs. The four different scores are detailed below:

1. Excluding 20 GWAS index SNPs and excluding the *APOE* region (Chr 19: Base Position (BP) 44.4Mb-46.5Mb)

2. Including all SNPs

3. Excluding 20 GWAS index SNPs, the *APOE* region and 10,697 heterogeneous SNPs in the IGAP data

4. Excluding 10,697 heterogeneous SNPs in the IGAP data

The tables show the prediction values for a number of different models. The first is a model just containing the number of *APOE* $\epsilon 4$ alleles. The second shows the additional prediction accuracy provided by the number of *APOE* $\epsilon 2$ alleles and the presense of the 20 GWAS SNPs, details of the index SNPs for these 20 GWAS hits are seen in Supplementary Table 11.2. The remaining models include the addition of POLARIS in the model, at a number of different p-value thresholds for SNP inclusion.

The AUC for the intelligently pruned GERAD data using PRS has been shown to be 0.782. This is 90% of the maximum prediction accuracy possible from genetic information only [38].

From Table 8.1 the maximum prediction accuracy (AUC=0.759) includes POLARIS in the model, using a p-value threshold of 0.4. A substantial proportion of the prediction accuracy is provided by the number of *APOE* $\epsilon 4$ alleles. There is some improvement using the addition of the number of *APOE* $\epsilon 2$ alleles and the 20 GWAS index SNPs. There is then also an increase in prediction accuracy when POLARIS is included in the model, suggesting the polygenic impact of SNPs which do not reach genome-wide significance. When comparing genetics alone, the maximum AUC for POLARIS is 0.744 compared to the AUC of PRS in pruned GERAD data of 0.745. There is a smaller difference between AUC when considering genetics alone.

Table 8.1: POLARIS Prediction Modelling, Excluding 20 GWAS hits and *APOE* region

Model	Sensitivity	Specificity	AUC	AUC (+age+sex)
$\epsilon4$	0.593	0.746	0.678	0.713
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs	0.665	0.665	0.714	0.732
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.0001)	0.667	0.667	0.716	0.734
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.001)	0.668	0.668	0.719	0.737
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.01)	0.672	0.672	0.727	0.743
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.05)	0.677	0.677	0.736	0.751
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.1)	0.677	0.677	0.739	0.755
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.2)	0.682	0.682	0.743	0.758
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.3)	0.682	0.682	0.743	0.758
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.4)	0.680	0.680	0.744	0.759
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.5)	0.680	0.680	0.743	0.758

The maximum AUC seen in Table 8.2 is 0.760, again, this is when POLARIS is included where the p-value threshold is 0.3. The maximum AUC in pruned GERAD data using PRS was 0.782 when PRS using a p-value threshold of 0.5 is included in the model [38]. The maximum AUC in POLARIS when ignoring age and sex is equivalent to the AUC found using PRS [38], although this is attained at a lower p-value threshold in POLARIS compared to PRS (0.3 and 0.5 respectively).

Table 8.2: POLARIS Prediction Modelling, Including 20 GWAS hits and *APOE* region

Model	Sensitivity	Specificity	AUC	AUC (+age+sex)
$\epsilon4$	0.593	0.746	0.678	0.713
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs	0.665	0.665	0.714	0.732
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.0001)	0.667	0.667	0.718	0.735
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.001)	0.665	0.665	0.721	0.738
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.01)	0.672	0.672	0.729	0.745
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.05)	0.678	0.678	0.737	0.752
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.1)	0.678	0.678	0.741	0.756
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.2)	0.683	0.683	0.745	0.760
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.3)	0.684	0.684	0.745	0.760
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.4)	0.680	0.680	0.745	0.760
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.5)	0.681	0.681	0.744	0.759
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.6)	0.677	0.677	0.742	0.757
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.7)	0.680	0.680	0.740	0.755
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.8)	0.682	0.681	0.738	0.753
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.9)	0.680	0.680	0.736	0.751
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<1)	0.677	0.677	0.732	0.748

Table 8.3 shows the maximum prediction accuracy to be 0.765, when POLARIS considering only SNPs with a p-value less than 0.2 is included. When considering the prediction accuracy of models including genetic information only, POLARIS attains a higher maximum AUC of 0.750 compared to that of 0.745 using PRS [38].

Table 8.3: POLARIS Prediction Modelling, Excluding 20 GWAS hits, *APOE* region and 10,697 heterogeneous IGAP SNPs

Model	Sensitivity	Specificity	AUC	AUC (+age+sex)
$\epsilon4$	0.593	0.746	0.678	0.713
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs	0.665	0.665	0.714	0.732
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.0001)	0.670	0.670	0.716	0.734
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.001)	0.670	0.671	0.721	0.738
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.01)	0.674	0.674	0.730	0.745
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.05)	0.680	0.680	0.740	0.755
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.1)	0.679	0.679	0.745	0.760
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.2)	0.686	0.687	0.750	0.765
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.3)	0.685	0.685	0.750	0.764
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.4)	0.685	0.685	0.750	0.764
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.5)	0.684	0.683	0.748	0.763

Table 8.4 shows the maximum prediction accuracy across all of the different POLARIS scores constructed. The AUC is 0.766 when a POLARIS score including SNPs with p-values less than 0.2 is included. This value is slightly lower than the AUC using PRS, this is likely due to the inflation in PRS caused by very small LD which is not removed by LD pruning. It is also noteworthy that POLARIS attains maximum prediction at a smaller p-value threshold (0.2 compared to 0.5). Again when ignoring age and sex in the prediction model, POLARIS attains a higher maximum AUC of 0.751 when a p-value threshold of 0.2 is used, compared to 0.745 found using PRS in pruned data.

Table 8.4: POLARIS Prediction Modelling, Including 20 GWAS hits and *APOE* region and Excluding 10,697 heterogeneous IGAP SNPs

Model	Sensitivity	Specificity	AUC	AUC (+age+sex)
$\epsilon4$	0.593	0.746	0.678	0.713
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs	0.665	0.665	0.714	0.732
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.0001)	0.663	0.663	0.717	0.734
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.001)	0.667	0.667	0.722	0.739
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.01)	0.673	0.672	0.731	0.747
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.05)	0.682	0.682	0.742	0.757
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.1)	0.683	0.683	0.746	0.761
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.2)	0.686	0.686	0.751	0.766
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.3)	0.685	0.685	0.751	0.765
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.4)	0.688	0.688	0.750	0.765
$\epsilon4$ + $\epsilon2$ + 20 GWAS SNPs + POLARIS(p<0.5)	0.682	0.682	0.749	0.764

Table 8.5 displays the AUC for all three methods; POLARIS, PRS and LDpred. The PRS AUC is taken from Escott-Price et al. (2015) [38]. LDpred requires the proportion of causal SNPs, cf, to be defined; prediction results are presented for cf=0.0001 and 1. The LDpred prediction results are higher when all SNPs are assumed to be causal. All approaches give similar results, with POLARIS having the highest prediction for genetics

only (excluding age and sex) and PRS having the highest prediction when including age and sex.

Table 8.5: Prediction Comparison Between POLARIS, PRS and LDpred; Excluding 20 GWAS hits, *APOE* region and 10,697 heterogeneous IGAP SNPs

	POLARIS		PRS		LDpred, cf=1		LDpred, cf=0.0001	
Model	AUC	AUC (+age+sex)	AUC	AUC (+age+sex)	AUC	AUC (+age+sex)	AUC	AUC (+age+sex)
α + RS(p<0.0001)	0.716	0.734	0.717	-	0.716	0.734	0.717	0.734
α + RS(p<0.01)	0.730	0.745	0.729	-	0.724	0.740	0.717	0.735
α + RS(p<0.05)	0.740	0.755	0.738	-	0.733	0.748	0.717	0.735
α + RS(p<0.1)	0.745	0.760	0.740	-	0.736	0.752	0.717	0.735
α + RS(p<0.5)	0.748	0.763	0.745	0.782	0.738	0.754	0.717	0.735

where cf represents the proportion of SNPs assumed to be causal in LDpred and $\alpha = \epsilon 4 + \epsilon 2 + 20$ GWAS SNPs.

8.4 Discussion

This chapter presents the use of POLARIS for all SNPs across the whole genome. POLARIS combines the advantages of PRS and spectral analysis of the genetic data. The motivation for this is to predict case/control status or the trait liability captured by all SNPs, this will have more case/control discriminatory power than set-specific prediction. POLARIS has the advantage in this whole genome approach that data does not require LD pruning prior to analysis, as the LD correction occurs as part of the POLARIS approach.

LDpred is an alternative Bayesian method which adjusts effect sizes (βs) for LD between SNPs. LDpred has an infinitesimal model which assumes that all SNPs are causal, and the noninfinitesimal model does not. The noninfinitesimal model uses a p-value threshold for the inclusion of SNPs, similar to the p-value inclusion thresholds used in POLARIS and PRS.

In simulated data, POLARIS and LDpred show equivalent power for all sample sizes considered, but POLARIS has slightly higher power, except when test and discovery sets have N=10,000, although LDpred has mildly inflated type I error in this case.

The β adjustment for LD between SNPs was also compared between POLARIS and LDpred; both methods show little to no adjustment when there is no LD between SNPs,

but LDpred shows some unusual results when the test set used to estimate LD is larger, with very small adjustment in the case of moderate LD. This opposes the POLARIS adjustment which shows increasing adjustment with increasing LD between SNPs, as expected.

POLARIS is shown to be a powerful method in application to the whole genome compared to LDpred in the real AD data. POLARIS has a larger maximum $-log_{10}$(p-value) compared to both LDpred and unadjusted PRS. Power for the POLARIS method is increased by thresholding the SNPs included in the score, although this then decreases when the p-value threshold is greater than 0.5, suggesting the sensitivity of POLARIS to the signal-to-noise ratio. This is demonstrated in AD data, see Figure 8.6, where the strength of association with AD is increased as the inclusion threshold increases; the additional information contained in SNPs which are not individually associated are incorporated into the score. Of course, if POLARIS is used for a sequence of inclusion thresholds, then a multiple testing correction of the results will need to be applied, as for standard PRS.

POLARIS is computationally expensive when the score is computed for a large number of SNPs. This is because matrix inversion of a large correlation matrix requires huge amounts of RAM and computational time. A set-based POLARIS software written in Python is currently available, although it is only able to compute the POLARIS score for a maximum of 200 SNPs in a set, therefore, this software is not able to be used for whole genome analysis, particularly with the increasing use of imputed data with a larger number of SNPs. An updated software written in C is currently under development which will enable the production of a POLARIS score for a larger number of SNPs, by utilising packages created to optimise matrix inversion. Ideally, the software will be able to invert correlation matrices of at least 500,000 SNPs.

The POLARIS score is able to well predict whether a person has AD or not (AUC including age and sex=0.766), this is almost as high as the reported AUC for PRS which is 0.782. When considering genetics only, POLARIS has higher prediction accuracy (AUC=0.751) compared to that found using PRS in pruned data (AUC=0.745). POLARIS has higher prediction accuracy based on genetics alone, because the data does not require LD pruning

and thus more SNPs are included in the score. When age and sex are additionally included in the model to assess prediction accuracy, PRS is better able to predict case/control status compared to POLARIS. This could be explained by the additional SNPs included in the POLARIS score potentially being linked to aging, so perhaps the effect of age is partially explained by these SNPs.

Prediction accuracy is increased above *APOE* and GWAS signals by including the POLARIS score, suggesting the polygenicity of AD, since SNPs which do not reach genome-wide significance still contribute to the ability to differentiate between cases and controls.

The main difference between POLARIS and LDpred is that POLARIS uses the square inverse of the correlation matrix, whereas LDpred uses the inverse. In essense, LDpred adjusts the genotypes and the βs for LD, but POLARIS adjusts the genotypes only. In POLARIS, it is not necessary to adjust the βs as well since these are calculated for each SNP separately, so are not influenced by LD. This approach is more similar to using the PRS method in pruned data, since LD is only removed from the test set, and βs remain unadjusted. POLARIS uses the correlation matrix to adjust the test set, assuming that correlation seen is LD between the SNPs rather than a correlation caused by SNPs which are associated with disease. Figure 8.7 shows the correlation caused by associated SNPs, this is done by running 1,000 simulations generating two independent SNPs which were both associated with disease in 10,000 individuals. The strength of each SNP's association to disease was varied from OR = 1 to OR = 4. Both SNPs had a MAF of either 0.2 or 0.3. The overall correlation, and the correlation in cases and controls separately was computed for each simulation. The mean of these values was found across all 1,000 simulations. In cases only and controls only the correlation coefficients are approximately zero, suggesting that SNPs are independent. However, there is a larger correlation in all the data, with correlation coefficient r increasing as the strength of association and MAF increase. LDpred applies the assumption of no association between SNPs and disease in the test set to the discovery data which is not appropriate when SNP effect sizes in the discovery set are known. This may explain the larger maximum $-log_{10}$(p-value) in POLARIS compared to LDpred in real AD data, since it does not over adjust for LD.

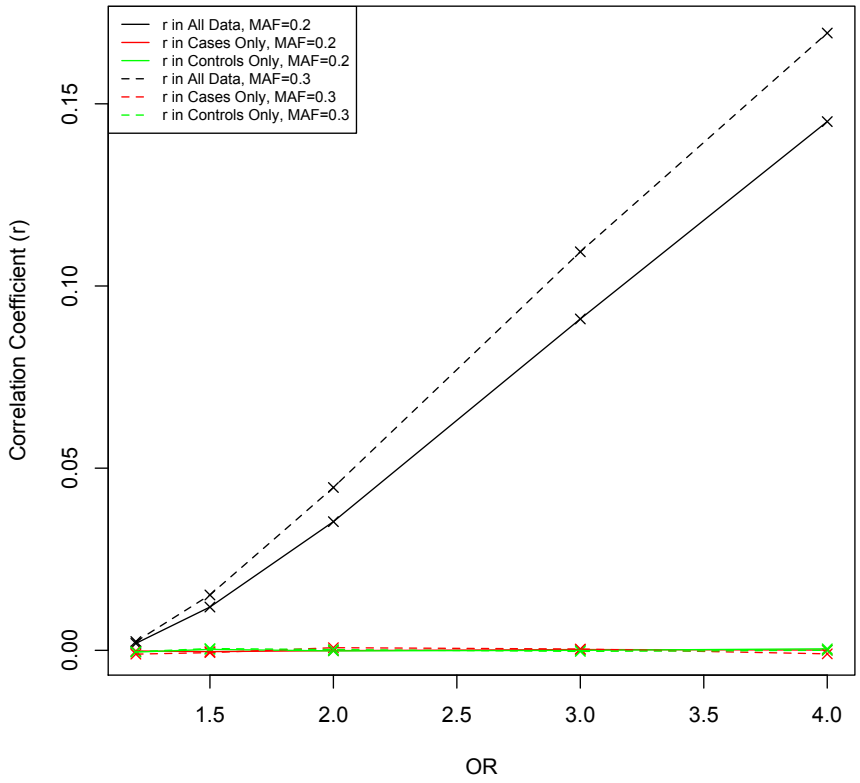


Figure 8.7: Plot of Correlation Coefficient, r, for Varying ORs. In All Data (Black) and Cases (Red) and Controls (Green) Separately for MAF=0.2 (solid) and MAF=0.3 (dashed)

PRS has consistently lower p-values compared to multivariable regression with all SNPs in the PRS; this is because the SNPs contributing to the PRS are prioritised for association with the disease, so the risk alleles will be more common among cases for each SNP. Therefore, even if associated SNPs are pruned for LD, they appear to be correlated because they are associated with disease. In POLARIS, all SNPs are included in the score, and so there is not as high a proportion of associated SNPs and this effect may not be seen. Although it depends on the SNPs which are selected, if those on the genotype chip were prioritised to be associated with disease, then perhaps this same effect would be seen.

9 Application of POLARIS as Cross Disorder Analysis

9.1 Introduction

GWAS studies have helped to determine the genetic architecture of a number of complex disorders. A larger amount of heritability is explained using PRS over single SNP effects found from GWAS. PRS [37] incorporates the effect of a large number of SNPs across the whole genome, not all of which individually reach genome-wide significance [131].

PRS can also be used to find commonality between different disorders. Brain disorders display comorbidity and often share symptomatic manifestations, suggesting potential biological overlap; for example, depression and anxiety have been shown to overlap [131]. The summary statistics from a related disorder can be utilised to incorporate additional information [132].

It is possible to identify disease aetiology or shared clusters of phenotypes by examining the genetic correlation between disorders, since this may identify shared genetic liability [133]. LD Score regression [134] has the functionality to compute cross-disorder genetic correlations using GWAS summary statistics and is not affected by sample overlap [135].

Like PRS, POLARIS can be used as a gene-based analysis using AD GERAD data and the effect sizes (weights) from other disorders. This chapter demonstrates this potential application of POLARIS, whereas other chapters consider single traits only. Genes which reach gene-wide significance from this analysis can potentially display commonality between AD

and other disorders. Significant genes may implicate particular disease mechanisms which are shared by the two disorders. A profile of significant genes across a number of different disorders may aid in a better understanding of the aetiology of AD. It may also be possible to differentiate different disorders since the gene effect sizes can be compared across different disorders.

Most GWAS studies are required to make GWAS summary statistics publically available. The analysis which we suggest in this chapter focuses predominantly on other brain disorders; of which a wide selection are available on the Psychiatric Genetics Consortium (PGC) website [35], for example SZ, Bipolar disorder (BIP), Major depressive disorder (MDD). In addition, the cross disorder gene analysis also investigates commonality between AD and Parkinson's Disease (PD) [136] and AD and Coronary Artery Disease (CAD) [137].

9.1.1 Objectives

This Chapter aims to;

- Demonstrate the application of POLARIS as a cross-disorder analysis.
- Investigate the genetic correlation between AD and a number of psychiatric disorders, PD and CAD. This will determine any potential biological overlap between disorders.
- Compute gene-based POLARIS scores for AD in GERAD data using all different disorders as weights for the score. This may determine genes in common between disorders which may explain the biological mechanisms underlying the disorders.

9.2 Materials and Methods

POLARIS has been demonstrated as a single-trait method, but it is possible to use POLARIS as a multi-trait or cross-disorder analysis. Cross disorder gene-based analyses can be used to determine genes which are in common between AD and a number of different disorders. This may enable the identification of potential disease mechanisms which

explain how AD develops, and which mechanisms are in common with other disorders.

The disorders which have had summary statistic data downloaded, and the number of SNPs in common between both genotyped and imputed GERAD data (Ncases=3,332, Ncontrols=9,832) are seen in Table 9.1. There are multiple versions for a number of disorders; all versions are considered in this analysis. A large number of GWAS studies in a number of disorders use controls from the WTCCC, so these were removed from the GERAD data (N=168 were removed) in order to maintain independence between the two datasets.

Table 9.1: Information on Disorders with Data Downloaded

Disorder	Year	No. Individuals	No. SNPs	No. Common SNPs with genotyped GERAD	No. Common SNPs with imputed GERAD
Anxiety	2016	17,310	6,330,995	409,824	2,665,838
Major depressive disorder	2015	10,640	6,208,598	347,966	1,121,378
Major depressive disorder	2012	18,759	1,235,109	377,064	530,406
Bipolar disorder	2012	16,731	2,427,220	347,602	975,863
Schizophrenia	2012	51,695	1,252,901	342,789	510,050
Parkinson's disease	2014	17,352	7,734,960	415,137	2,158,500
Autism spectrum disorder	2015		9,499,589	417,196	3,243,956
Schizophrenia	2014	150,064	9,444,230	414,604	3,240,793
Attention deficit hyperactivity disorder	2013	5,422	1,230,535	352,782	503,667
Autism spectrum disorder	2013	10,263	1,168,835	386,458	538,616
Bipolar disorder	2013	11,810	1,233,533	330,116	490,280
Major depressive disorder	2013	16,610	1,232,793	377,777	524,588
Schizophrenia	2013	17,115	1,237,958	340,112	502,349
Neuroticism	2015	137,178	9,181,138	412,737	5,979,941
Bipolar disorder		51,710	13,414,632	417,618	3,321,367
Epilepsy	2014	34,853	5,968,967	312,344	2,807,803
Schizophrenia		35,976	8,064,799	409,803	1,234,241
Schizophrenia (info > 0.9)		35,976	5,471,113	360,141	1,103,578
Coronary artery disease	2017	63,731	9,026,567	416,003	3,080,078

9.2.1 Genetic Correlation

LD Score [134] is a software which is able to compute the genetic correlation [135] between disorders by LD score regression, using only GWAS summary statistics. This method uses the fact that SNPs in high LD will have higher χ^2 statistics compared to SNPs in low LD when considering polygenic traits. This is applicable to multiple disorders by using the product of z-scores from two different datasets. LD score regression does not require the two datasets to be independent, since overlap between the datasets only effects the intercept in the model rather than the genetic correlation estimate [135].

The genetic correlation was computed between AD data (both GERAD genotype and

IGAP summary statistics) and all other disorders. These genetic correlations were plotted into a heatmap using the corrplot package in R software. The heatmap displays both the direction and strength of the correlation and the associated p-value, indicating the statistical significance of the genetic correlation. The genetic correlation is presented to indicate potential disorders which may have common underlying mechanisms, disorders which are highly correlated may have more genes in common found from the POLARIS gene-based analysis.

9.2.2 Cross Disorder POLARIS

POLARIS gene-based method [96] is a powerful method which extends upon the advantages of PRS [37] by additionally adjusting for LD between SNPs. The SNP weightings used in the score may be from a different trait than that in the test data. The LD adjustment in POLARIS uses the spectral decomposition of the correlation matrix to adjust AD SNP genotypes and replace them with LD adjusted dosages. This produces a risk score for AD weighted with SNP effect sizes from another disorder.

Merged SNPs between the test and discovery sets were assigned to genes using GENCODE (v19) gene models [78]. Only genes with known gene status and those marked as protein coding were used. There was no window used around the gene, only SNPs within the start and end position of the gene were included.

POLARIS scores are calculated per subject per gene and the overall association of the gene is determined using logistic regression of AD status on the LD adjusted polygenic risk score and additional population covariates. POLARIS is utilized in a number of cross disorder analyses to determine any commonality between AD and psychiatric disorders, PD or CAD. This is done by training on psychiatric, PD or CAD data and testing in AD GERAD data. Publically available GWAS data from the PGC [35] (`http://www.med.unc.edu/pgc/results-and-downloads`), data from the International Parkinsons Disease Genomics Consortium [136] and UK Biobank data for CAD [137] were used to determine overlap between AD and a number of disorders at a gene-based level. The full list of disorders investigated is seen in Table 9.1.

The cross disorder analysis initially considered the AD GERAD genotype data, and was then repeated in the GERAD imputed data, since this is likely more powerful as it contains a larger number of SNPs.

9.3 Results

9.3.1 Genetic Correlation

The genetic correlation between all disorders and AD as computed from LD Score regression [134][135] are shown in the heatmap in Figure 9.1. The heatmap shows the pairwise correlation between all disorders. The size of the block represents the p-value, with the largest square being a p-value of 0 and no square representing a p-value of 1. The colour of the square represents the genetic correlation coefficient, red shows a positive correlation and blue shows a negative correlation, these fade to white for a correlation of zero.

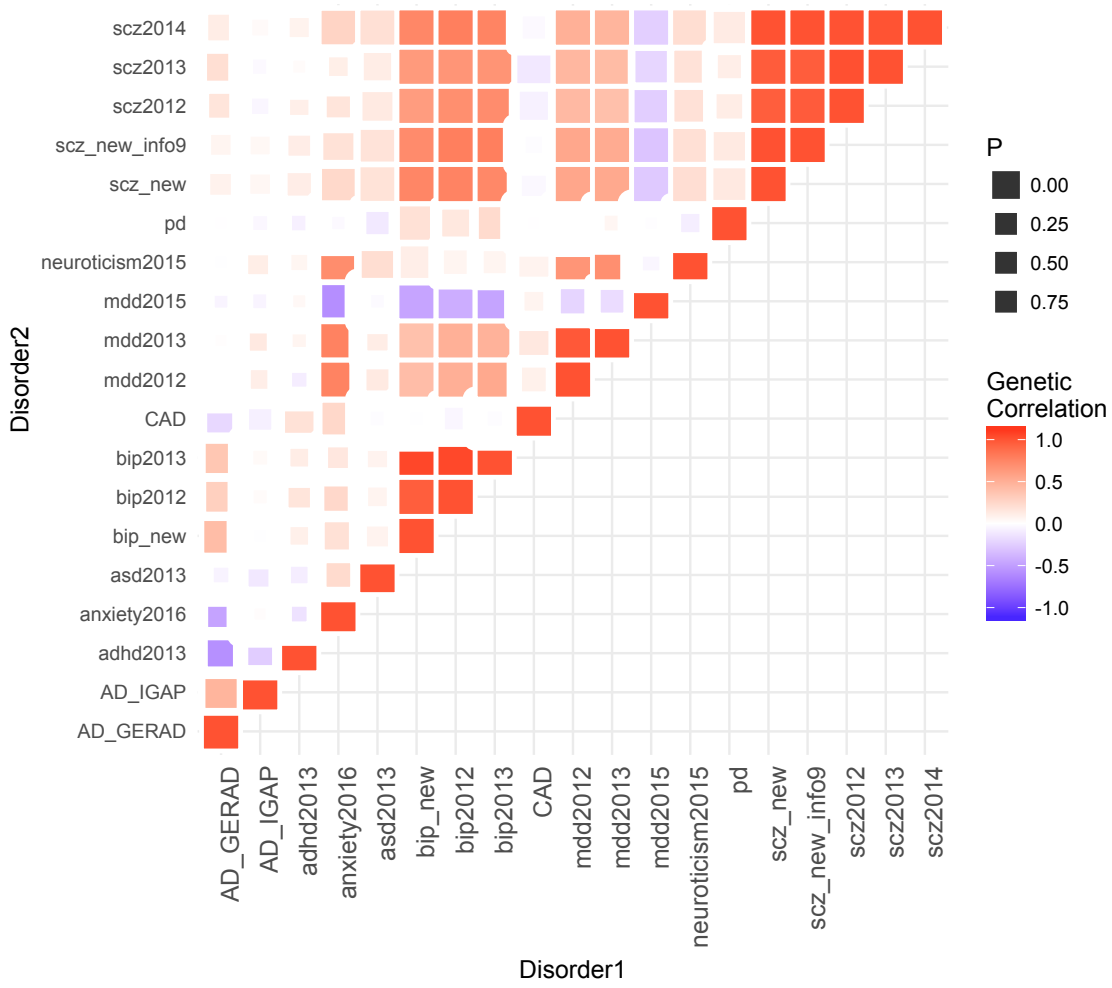


Figure 9.1: Plot of the Genetic Correlation Between All Disorders

There are multiple versions of the same data, for example SZ, from Figure 9.1 it is seen that these groups of studies which represent the same trait have high positive correlation. There is one exception to this; mdd2015 is negatively correlated with the other major depressive disorder studies. The two AD datasets, GERAD and IGAP show a strongly associated positive correlation. Of course, GERAD is a subset of the IGAP data, however, the genetic correlation estimate is not effected by this overlap, as only the intercept is influenced by this overlap in LD score regression [135]. The strong genetic correlation between the GERAD and IGAP data is likely due to both consortia containing GWAS data for AD and hopefully share similar SNPs.

The genetic correlation between AD and other disorders does not have particularly small p-values, with the lowest being 0.001 with bipolar disorder (bip_new). AD shows no genetic correlation with PD [138], MDD or neuroticism. AD displays a positive correlation with SZ and BIP and a negative correlation with Attention deficit hyperactivity disorder (ADHD), anxiety, Autism spectrum disorder (ASD) and CAD.

9.3.2 Cross Disorder POLARIS

For each of the disorders listed in Table 9.1, a table for the gene-wide significant genes for the multi-trait analysis between AD and each disorder are presented. The tables show all genes which reach gene-wide significance ($p < 2.5 \times 10^{-6}$), and genes with suggestive significance ($2.5 \times 10^{-6} < p < 0.00001$) are shown in italics. For each disorder, a table for both genotyped and imputed data are presented, unless no at least suggestively significant genes were determined.

The *APOE* locus is present in all analyses, although it is not as strongly associated as in AD PRS. The presence of *APOE* in the cross-disorder analyses is due to the very large effect in the AD data, rather than a commonality between AD and other disorders. The effect is so large in the AD data, that even with a SNP weighting of approximately zero from the other disorder, the effect of *APOE* remains. Therefore, the *APOE* locus is removed from all analyses, any genes on chromosome 19 between base positions 44.4Mb and 46.5Mb. Similarly, *CLU* is found to be associated for almost all of the cross-disorder analyses, this is again influenced by the strong effect in AD rather than any commonality between AD and other disorders, so is not presented in the results.

9.3.2.1 Attention Deficit Hyperactivity Disorder (ADHD)

The gene-based results from the cross disorder analysis in AD weighted with ADHD effect sizes is seen in Table 9.2. In the GERAD genotype data, there is suggestive evidence that *IL1RL1* has an association with AD supported by ADHD; this gene is on chromosome 2 and is an interleukin receptor which is part of the immune response [139]. In fact, *IL1RL1*

has been identified in an immunity pathway for psychiatric disorders (SZ, BIP and MDD) [76].

Table 9.2: Results for AD Gene-Based Analysis Using ADHD Summary Statistics as Weights

GENE	CHR	NoSNPs	P
IL1RL1	2	4	8.9×10^{-6}

(i) GERAD genotype data

GENE	CHR	NoSNPs	P
IL1RL1	2	145	4.9×10^{-6}

(ii) GERAD imputed data

In the imputed GERAD data, again *IL1RL1* shows a suggestive commonality between AD and ADHD.

9.3.2.2 Anxiety

The gene-based cross-disorder analysis using anxiety weights in AD GERAD genotype data and imputed data only determine genes within the *APOE* locus or *CLU*.

9.3.2.3 Autism Spectrum Disorder (ASD)

The gene-based results for genes in common between AD and ASD are shown in Table 9.3. In the genotype data, only genes in the *APOE* locus or *CLU* are found to be gene-wide significant.

Table 9.3: Results for AD Gene-Based Analysis Using ASD Summary Statistics as Weights

Data Version	GENE	CHR	NoSNPs	P
2015	*BSND*	1	14	8.5×10^{-6}

(ii) GERAD imputed data

In the GERAD imputed data, with the 2015 version of the ASD data, the *BSND* gene which resides on chromosome 1 shows a suggestive association.

9.3.2.4 Bipolar Disorder (BIP)

The gene-based results for AD GERAD genotype data weighted with BIP effect sizes are seen in Table 9.4. In all versions of the BIP data, only *CLU* or genes in the *APOE* region show any association.

Table 9.4: Results for AD Gene-Based Analysis Using BIP Summary Statistics as Weights

Data Version	GENE	CHR	NoSNPs	P
Latest	CSDC2	22	26	1.6×10^{-6}

(ii) GERAD imputed data

In the GERAD imputed data a large number of genes in the *APOE* locus are gene-wide significant (not presented). In addition, *CSDC2* on chromosome 22 is in common between BIP and AD.

9.3.2.5 Coronary Artery Disease (CAD)

The AD gene-based analysis using CAD weightings only determines association with genes in the *APOE* locus or *CLU* in both the genotype and imputed data.

9.3.2.6 Major Depressive Disorder (MDD)

Aside from genes in the *APOE* region, there are no genes which reach even a suggestive significance of commonality between AD and MDD in the genotype or imputed data.

9.3.2.7 Neuroticism

Table 9.5 shows the results for the gene-based analysis in AD using neuroticism effect sizes as weights. This analysis in the genotype data found no gene-wide significant genes other than those in the *APOE* locus.

Table 9.5: Results for AD Gene-Based Analysis Using Neuroticism Summary Statistics as Weights

GENE	CHR	NoSNPs	P
ZNF525	19	57	1.1×10^{-6}

(ii) GERAD imputed data

The cross disorder analysis using the imputed GERAD data as a test set also only finds an association with a gene which is likely caused by the large effect of *APOE*; *ZNF525*, but does not reside within the defined *APOE* locus.

9.3.2.8 Parkinson's Disease (PD)

Table 9.6 shows the results from the gene-based analysis in AD with POLARIS weights from PD. In the GERAD genotype data, *ZNF525* shows a suggestive association, this gene is on chromosome 19, approximately 7Mb upstream of the *APOE* region, so this association may be caused by *APOE*.

Table 9.6: Results for AD Gene-Based Analysis Using PD Summary Statistics as Weights

GENE	CHR	NoSNPs	P
ZNF525	19	5	3.3×10^{-6}

(i) GERAD genotype data

GENE	CHR	NoSNPs	P
ZNF525	19	57	2.7×10^{-6}

(ii) GERAD imputed data

The GERAD imputed data also only finds a suggestive association with the *ZNF525* gene which is on chromosome 19. Again, this is close to the *APOE* locus so it may be affected by *APOE*.

9.3.2.9 Schizophrenia (SZ)

The results for the gene-based analysis in AD weighted with SZ effect sizes are seen in Table 9.7. For all versions of the SZ data, no gene-wide or suggestively associated genes are found in common between genotype AD data and SZ.

Table 9.7: Results for AD Gene-Based Analysis Using SZ Summary Statistics as Weights

Data Version	GENE	CHR	NoSNPs	P
2014	*BSND*	1	14	6.9×10^{-6}
Latest	*TMEM61*	1	28	9.3×10^{-6}

(ii) GERAD imputed data

The cross disorder analysis carried out in the imputed GERAD data finds two additional suggestively significant genes across all versions of the SZ data. Both suggestive genes reside on chromosome 1, these are *BSND* and *TMEM61*.

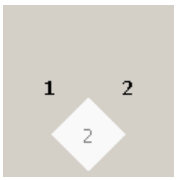
9.4 Discussion

This Chapter demonstrates the application of POLARIS as a multi-trait or cross disorder analysis, where the effect sizes used to weight the score are for a disorder which differs to the phenotype used in the logistic regression model.

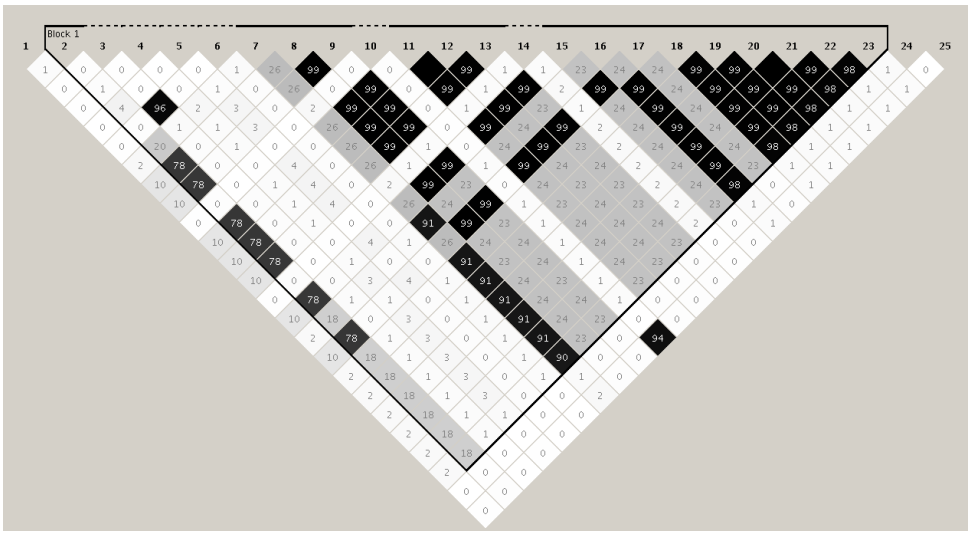
A few gene-wide or suggestively significant genes are identified from this cross-disorder analysis, the biological implication of these genes requires further investigation. The effect of *APOE* in AD data is still apparent, despite informing the POLARIS score with effect sizes from other disorders, *APOE* was detected in all analyses with $p < 10^{-20}$. Clearly, the effect of *APOE* in AD dominates the effects of other disorders, but this does not indicate commonality between genes in the *APOE* locus, therefore, these results should be interpreted with care. This is also seen for the *CLU* gene which is strongly associated with AD, and is therefore present in the majority of cross-disorder analyses, however, this does not suggest that *CLU* is associated with both disorders, but rather that the effect of *CLU* in AD dominates the other disorder. Therefore, genes which have a strong effect in one disorder in a cross-disorder analysis should always be carefully assessed.

This analysis has been carried out in both genotype and imputed data, the use of imputed AD data adds strength to this analysis and results in stronger and additional associations. However, in the case of *CLU*, in a number of cases the gene-wide significance seen in genotype data disappears when imputed data is used, this may be due to the effect of

additional SNPs cancelling the effect of SNPs in the genotype data only. In the genotype data, only two SNPs are included in *CLU* which have almost no LD with one another. In the imputed data, 25 SNPs are included in *CLU*, some of these SNPs show very high LD with one another, see Figure 9.2 for the LD structure of SNPs in *CLU*. This differing LD structure may explain why associations with *CLU* are apparent in the genotype data but not in the imputed data.



(a) Genotype Data



(b) Imputed Data

Figure 9.2: LD Plot for *CLU* SNPs

Cross-disorder gene-based analyses may find genes which contribute to common clinical outcomes/symptoms and therefore treatments could be directed to these symptoms.

Predominantly only genes which are within the *APOE* locus or have been previously identified by AD GWAS, e.g. *CLU*, and as such only represent the association in AD have been determined. Some additional suggestive or non-GWAS genes have been found, such as *IL1RL1* and *CSDC2*, although further research is necessary to assess whether these genes have credible biological implications in AD. The gene-wide significance threshold used in this analysis ($p < 2.5 \times 10^{-6}$) does not take account of the multiple disorders

tested, so these additional genes would likely not withstand the additional multiple testing correction.

Understanding the relationship between two disorders or phenotypes can lead to the inference of GWAS summary statistics based on a function of disorders/phenotypes which each have GWAS summary statistics available. This approach is termed GWIS; Genome-wide Inferred Study [140], and would be particularly applicable to psychiatric disorders which have a substantial shared phenotype, for example, cognition in schizophrenia and bipolar disorder. This approach may be applicable in that summary statistics for disorders which are related can be inferred and used in the POLARIS score as weights.

The causes of genetic correlation can not be investigated using the POLARIS method. A novel Genomic Structural Equation Modelling (SEM) [133] method has been developed which enables the comparison of a number of multivariate genetic architectures. This would be a suitable extension to this analysis to further investigate significant findings.

Additionally, the results from cross-disorder analyses can be used to improve disease prediction. For genetically correlated disorders, it is possible to increase the disease risk prediction accuracy by including information from the correlated disorder. The approach to do this is called weighted multi-trait summary statistic best linear unbiased predictor (wMT-SBLUP) [141].

10 Discussion and Implications

10.1 Conclusions

The gold standard of medical treatment is to develop personalised medicines, the best way to approach this is using genetics. Complex traits such as AD have been shown to have a large number of genes associated with them. These have been identified using the GWAS method, which require large sample sizes to detect the small individual effect of SNPs. In order to determine novel variants associated with AD, it is necessary to use methods which gain power compared to GWAS.

The primary aim of this thesis was to extend methods beyond GWAS in order to identify novel genes associated with AD. The POLARIS method has been introduced and is a powerful extension to the PRS approach which takes account of LD between SNPs. Therefore, data is not required to be pruned to remove LD and thus more information is able to be incorporated into the risk score. POLARIS aggregates the small individual SNP effects for all SNPs within a set into a larger polygenic component, whilst weighting with SNP effect sizes from an independent dataset. This POLARIS approach can be applied to any set of SNPs such as genes, pathways and the whole genome.

The POLARIS methodology was applied to the latest imputation of the GERAD data for a gene-based analysis. Three novel genes associated with AD were determined; *PPARGC1A*, *RORA* and *ZNF423*. These novel genes all have credible biology related to AD. In addition, two novel genes, *CSMD1* and *MACROD2* were found using the PRS gene-based method, also in the HRC imputed GERAD data.

10.2 Discussion

The analyses in this thesis are carried out in very large (N=13,164) recently imputed GWAS data where individual genotype data are available, additionally, the largest available GWAS in AD containing summary statistics only is utilised to increase power. Recently available methods which have been shown to be more powerful than standard methods are used. The use of the best methods and large data enable this to be a valid study, and results found are reliable. However, replication of determined genes in an independent dataset would be desirable.

Set-based analysis has been shown to be a powerful approach compared to single SNP analyses. A set is considered to be any set of SNPs. The small effect of individual SNPs are combined into a larger aggregate effect, and therefore there is more power to be able to find an association. Gene-based analyses are more robust than single SNP analyses across different populations. For single SNP analyses, because of the LD between SNPs, different SNPs may be determined in different populations. Pathway or gene-set analyses combine genes which have a similar biological function. By determining these pathway associations with disease, it is possible to implicate biological mechanisms of disease. Gene-based analysis has been shown to be more powerful than single SNP analyses by using the gene-based analysis available in MAGMA software, additional independent gene associations are determined in identical data. The difficulty with a gene-based analysis is how to define the gene, whether just to include SNPs within the start and end base position of the gene or whether to include a window around the gene. Therefore, it was additionally investigated and shown that using a flanking region of 35kb upstream and 10kb downstream around the gene to contain transcriptional elements increases the power of a gene-based analysis.

PRS is a method to determine the polygenic component of disease from these GWAS signals, by combining the genotypes of SNPs and weighting them with SNP effect sizes from an independent dataset. PRS can be considered as a set-based analysis by only considering SNPs within the set in the risk score. This produces a risk score per person per set. The association of the set-based risk score and disease can then be assessed

using a regression model adjusting for population covariates. PRS as a set-based method was compared to MAGMA-PCA, MAGMA-SUMMARY, Simes' and Fisher's methods in simulated data. The PRS set-based method improves upon the power of MAGMA-PCA in the test set only and has particularly high power when a real LD structure is considered.

The PRS set-based method was then applied to the real AD GERAD data, using the IGAP-noGERAD summary statistics as weights. Both a gene-based and pathway analysis were considered. Two additional genes, *CSMD1* and *MACROD2*, have been found to be associated with AD from the PRS gene-based analysis in the imputed GERAD data. *CSMD1* has been implicated in AD, familial Parkinson's disease [92] and cognitive performance [93]. The *MACROD2* gene has been implicated in neurological disorders [94]. Pathway PRSs were produced for the pathways which have been previously identified as associated with AD; two of these pathway risk scores were associated with AD when adjusting for the baseline PRS association, these pathways were the immune response and hematopoietic cell lineage pathways. The PRS method has the advantage of producing a risk score per person per set, and this score can be used for further functional analyses, or prioritising subjects for clinical trials. The PRS approach is also able to improve power by using the SNP effect sizes from an independent dataset to weight the score. The main issue with this PRS method is that data must be pruned for LD prior to producing the score, this removes a large amount of data, in addition, the method requires two independent datasets, one of which must contain individual genotype data.

The POLARIS methodology was introduced to extend upon the PRS method by additionally adjusting for LD between SNPs so that data does not require LD pruning. POLARIS uses the spectral decomposition of the correlation matrix to produce LD adjusted dosages for each individual, these adjusted dosages are then used in the risk score. POLARIS was compared to both MAGMA-PCA and MAGMA-SUMMARY in both simulated and real data. POLARIS was shown to give correct type I error in all simulated data, and therefore is successfully adjusting for the LD between SNPs. The power of POLARIS lies between that of MAGMA-PCA in the test set only and MAGMA-PCA in the combined test and discovery set. In practice, researchers would use all available genotype data, and would only use PRS or POLARIS if an additional independent dataset containing sum-

mary statistics were available. A POLARIS software is available for set-based analyses, but this is only able to handle very small sets due to the computational demand of matrix inversion. An updated software is currently under development which aims to be able to handle larger sets and reduce computational time. POLARIS set-based approach was applied to the real GERAD AD data using IGAP-noGERAD summary statistics as weights. Three novel genes were found when the gene-based analysis was considered in imputed GERAD data using a gene window of 35kb upstream and 10kb downstream which incorporates additional transcriptional elements into the gene. The three novel genes found to be associated with AD are *PPARGC1A*, *RORA* and *ZNF423*, all of which have biological relevance to AD. *PPARGC1A* is linked to the generation of amyloid plaques and energy metabolism, *RORA* is differentially expressed in the hippocampus and *ZNF423* resides in an AD-specific protein network. POLARIS pathway scores were produced for the eight pathways previously found to be associated with AD. Four of these POLARIS pathways are associated after adjustment for the baseline level of association; these are the immune response, cholesterol transport, hematopoietic cell lineage and clathrin pathways. Like PRS, POLARIS produces a risk score per person per set which is useful for future analyses, and increases power by incorporating an additional external dataset and by not requiring LD pruning which utilises all available information.

PRS is most often used to determine the overall polygenicity of disease by incorporating all SNPs which show some association to disease, but do not reach genome-wide significance. Like PRS, POLARIS is also applicable to any set of SNPs, including all SNPs across the genome. POLARIS was compared to LDpred which is an alternative method which also adjusts for LD and computes a risk score. POLARIS computed across the whole genome in GERAD data using IGAP-noGERAD summary statistics as weights has a larger maximum $-log_{10}$(p-value) compared to both LDpred and PRS. Although, it is also seen that POLARIS is sensitive to the signal-to-noise ratio caused by the inclusion of SNPs which have high p-values. This sensitivity requires further investigation and the POLARIS methodology altered in order to remove this issue. The POLARIS score is able to well predict whether a person has AD or not (AUC including age and sex=0.766), this is almost as high as the reported AUC for PRS which is 0.782. When considering

genetics only, POLARIS has higher prediction accuracy (AUC=0.751) compared to that found using PRS in pruned data (AUC=0.745). POLARIS has higher prediction accuracy based on genetics alone, because the data does not require LD pruning and thus more SNPs are included in the score. When age and sex are additionally included in the model to assess prediction accuracy, PRS is better able to predict case/control status compared to POLARIS. This could be explained by the additional SNPs included in the score potentially being linked to aging, so perhaps the effect of age is partially explained by these SNPs. Prediction accuracy is increased above *APOE* and GWAS signals by including the POLARIS score, suggesting the polygenicity of AD, since SNPs which do not reach genome-wide significance still contribute to the ability to differentiate between cases and controls. This whole genome analysis should also be carried out in the imputed GERAD data, the inclusion of additional SNPs may improve the prediction accuracy of the POLARIS score. This analysis requires the development of the POLARIS C code which doesn't require the use of too much RAM or computational time. Whilst improving the POLARIS methodology, it should be considered whether it is possible to incorporate additional functional information into the score, by using a different weighting for the SNPs in the score, such as methylation or expression values.

PRS has consistently lower p-values compared to multivariable regression with all SNPs in the PRS; this is because the SNPs contributing to the PRS are prioritised for association with the disease, so the risk alleles will be more common among cases for each SNP. Therefore, even if associated SNPs are pruned for LD, they appear to be correlated because they are associated with disease. In POLARIS, all SNPs are included in the score, and so there is not as high a proportion of associated SNPs and this effect may not be seen. Although it depends on the SNPs which are selected, if those on the genotype chip were prioritised to be associated with disease, then perhaps this same effect would be seen.

The POLARIS method so far has been utilised using independent datasets for the same trait. The final analysis investigates the use of weighting the POLARIS gene-based score with effect sizes from alternative, potentially related disorders. This enables the identification of genes in common between disorders and may implicate potential biological mechanisms in AD. This gene-based analysis was weighted for a number of different psychiatric

disorders, CAD and PD. Some genes show at least a suggestive association, however, these associations will likely not withstand the additional multiple testing correction required due to the number of different disorders assessed. This work may also be further extended by comparing the POLARIS scores and score using the Genomic Structural Equation Modelling (SEM) [133] method which compares multivariate genetic architectures.

10.3 Limitations

For all genes determined in this thesis, it would be desirable to replicate these findings in an independent dataset, however, it is often difficult to get access to data which contains individual genotypes.

POLARIS is sensitive to signal-to-noise ratio; by using higher p-value inclusion thresholds, the overall POLARIS significance decreases due to the inclusion of a larger number of unassociated SNPs. A number of approaches were trialed in an attempt to remove this issue, but unfortunately none were successful. Given more time, this sensitivity would have been further investigated, it is hoped that using the pseudoinverse for the whole genome approach might resolve this issue.

The current POLARIS software available to other researchers is only appropriate as a set-based analysis, where the SNP sets do not exceed 200 SNPs. This is due to computational limitations; the amount of time required for the analysis and the amount of RAM necessary for the computation of the square inverse. An updated software is under development, but unfortunately it was not possible to complete this in time for the completion of this thesis. The new software uses packages which have been optimised for large matrix inversion, and will aim to enable researchers to use POLARIS as a whole genome approach.

Due to time limitations, the cross-disorder analysis was simply used to demonstrate the application of POLARIS as a multi-trait analysis but was not able to be extended to investigate alternative methods which are able to investigate the cause of genetic correlation [133] or use correlated traits to further improve prediction accuracy [141].

10.4 Further Work

Using the POLARIS and PRS gene-based analyses, five novel genes have been determined; these are *CSMD1*, *MACROD2*, *PPARGC1A*, *RORA* and *ZNF423*. Research has shown that these genes are biologically credible in AD, but further work is required to replicate these findings in independent data and functional studies are necessary to further assess the biological implications of these genes.

The POLARIS methodology requires further investigation to remove the signal-to-noise sensitivity observed in Chapter 8. This effect may be removed by using the pseudoinverse of the correlation matrix, since the whole genome analysis in this thesis does not use this approach. The ability to further test POLARIS would be aided by the development of an improved POLARIS software in C, which aims to reduce RAM and computational time requirements. Once this software is available, it would be possible to compute the whole genome POLARIS score in the GERAD imputed data which may be able to better predict whether or not a subject will develop AD.

In this thesis, the POLARIS methodology has only been used with a binary phenotype, however, quantitative traits may also be assessed by using a linear regression model. Currently, the set-based POLARIS software is only able to determine the association between POLARIS and binary traits, using a logistic regression model, however, the software will be updated to handle quantitative traits. Additionally, the software is only available as a self-contained test, this may also be updated to allow for a competitive test of association, which adjusts for the baseline level of association in the data.

For POLARIS there is also the possibility to use alternative weightings, such as functional SNP information like gene expression or methylation, instead of SNP effect sizes in alternative data. This may enable better AD risk prediction by the incorporation of relevant functional information.

It was noted in Chapter 8 that LDpred gives some strange β adjustments in the simulated data compared to POLARIS; investigation into the LDpred methodology does not fully explain the observed results, so additional research into this effect is necessary.

10.5 Implications

The production of a more accurate risk score using POLARIS means that subjects at high and low risk can be determined and used in clinical trials. Also, gene and pathway scores for each individual can be used to prioritise genes for functional studies to help aid in the understanding of the aetiology of AD. The inclusion of functional information into this score may enable an even more accurate prediction of disease.

Long term, the ability to accurately determine a person's disease risk which can be measured from birth may be hugely beneficial in the area of personalised medicine. It may lead to individuals being able to manage their health based on their disease risk, or for diseases to be treated prior to any symptom development. In addition, treatments can be tailored to a person's genetic information, which may aid in better response to treatments or reduced side effects using these precision medicines.

11 Supplementary Material

11.1 129 SNPs in Real Data Simulations

Table 11.1: Details of 129 SNPs Used in Real Data Simulations

CHR	SNP	BP	MAF	BETA	SE	P	HWE_P
1	rs11808115	50002165	0.0469	-0.1003	0.06783	0.1392	0.003202
1	rs4926816	50019849	0.3158	-0.03482	0.03075	0.2574	0.2761
1	rs1999875	50049490	0.1055	-0.0345	0.04616	0.4548	0.1277
1	rs11587574	50053885	0.08912	0.03151	0.04938	0.5234	0.7069
1	rs10888663	50081573	0.05323	-0.03818	0.06329	0.5463	0.1405
1	rs7540606	50086414	0.05318	-0.03739	0.0633	0.5548	0.1402
1	rs12741784	50088819	0.3748	-0.02467	0.02936	0.4008	0.8379
1	rs2051086	50105201	0.3233	-0.02594	0.03038	0.3933	0.9841
1	rs4408195	50128198	0.3641	-0.03757	0.02946	0.2023	0.5469
1	rs10489864	50161927	0.06659	0.07062	0.0558	0.2056	0.3614
1	rs11205641	50185075	0.4221	-0.03648	0.02878	0.205	0.7887
1	rs1474910	50192150	0.2233	-0.03369	0.03414	0.3237	0.5294
1	rs10788917	50227177	0.3644	-0.03866	0.02945	0.1894	0.5224
1	rs4926542	50263773	0.3155	-0.0201	0.03056	0.5107	0.9678
1	rs17105782	50286109	0.05269	-0.0489	0.0638	0.4434	0.19
1	rs17099766	50289163	0.04688	-0.105	0.06785	0.1218	0.00169
1	rs4926831	50290101	0.4228	-0.03158	0.02878	0.2726	0.7209
1	rs12128108	50293421	0.2249	-0.03435	0.03412	0.3141	1
1	rs4421632	50319197	0.09575	-0.07284	0.04922	0.1389	0.1445

1	rs10888672	50359987	0.3696	-0.02439	0.02931	0.4053	0.3999
1	rs12403905	50372032	0.3696	-0.02532	0.02932	0.3878	0.4321
1	rs9659092	50443589	0.4297	-0.03989	0.02874	0.1651	0.6311
1	rs4420126	50468739	0.3243	-0.02662	0.03037	0.3808	0.8893
1	rs4259707	50528591	0.2324	-0.0457	0.03374	0.1756	0.7881
1	rs6673246	50554373	0.05258	-0.0333	0.06363	0.6008	0.1885
1	rs4298778	50557867	0.04735	0.04141	0.06588	0.5297	0.3834
1	rs11583200	50559820	0.3822	-0.01865	0.02932	0.5247	0.3757
1	rs967582	50582172	0.3531	-0.006237	0.02957	0.833	0.4341
1	rs9436444	50589798	0.2278	-0.05812	0.03409	0.08823	0.9408
1	rs9436447	50590732	0.2279	-0.05847	0.0341	0.08639	0.9211
1	rs10493151	50598750	0.05269	-0.03855	0.06431	0.5489	0.3821
1	rs6588376	50602495	0.207	-0.06551	0.03552	0.06514	0.2322
1	rs3902720	50605337	0.3142	-0.07366	0.03078	0.01671	0.9355
1	rs17105974	50620945	0.05409	-0.06484	0.06341	0.3065	0.4424
1	rs2988269	50641944	0.3169	-0.07517	0.0307	0.01436	1
1	rs2841871	50684937	0.02739	-0.08059	0.08854	0.3627	0.8703
1	rs4926853	50709222	0.3324	0.0006043	0.0302	0.984	0.4318
1	rs17381266	50717206	0.09998	-0.03933	0.04794	0.412	0.146
1	rs3001644	50722989	0.08057	-0.05735	0.05286	0.278	0.5565
1	rs12030672	50735395	0.2452	-0.01705	0.03296	0.605	0.6546
1	rs12123613	50743010	0.07316	-0.02479	0.05452	0.6493	0.3036
1	rs1393637	50753599	0.03487	-0.00829	0.07768	0.915	0.5174
1	rs1875645	50789880	0.4546	0.002194	0.02834	0.9383	0.2836
1	rs6689749	50800516	0.4816	-0.01785	0.02825	0.5274	0.2421
1	rs4491095	50802855	0.3943	-0.01201	0.029	0.6788	0.8125
1	rs1393632	50814015	0.4608	-0.008859	0.02837	0.7549	0.5626
1	rs3862267	50822565	0.4603	-0.005301	0.02837	0.8518	0.5276
1	rs2765895	50822914	0.4846	-0.01975	0.02828	0.485	0.453
1	rs12071347	50839158	0.3002	-0.002649	0.0309	0.9317	0.7874

1	rs7525764	50844697	0.4878	0.003591	0.02823	0.8988	0.2873
1	rs3907676	50863202	0.3034	-0.002218	0.03078	0.9426	0.65
1	rs1278537	50882159	0.4736	0.01204	0.02817	0.6691	0.07458
1	rs12142848	50918388	0.0664	-0.02758	0.0571	0.6291	0.4817
1	rs3827730	50937848	0.3274	-0.05654	0.0303	0.06201	0.6069
1	rs12065210	50942784	0.09306	0.009107	0.04846	0.8509	0.2558
1	rs3789576	50945526	0.0901	0.001364	0.04928	0.9779	0.3383
1	rs10888690	50960521	0.4078	-0.004109	0.02865	0.8859	0.1005
1	rs7535374	50966897	0.1158	-0.04391	0.04501	0.3292	0.08108
1	rs1149795	50972564	0.1106	0.1015	0.0446	0.0229	0.3081
1	rs11581155	50976306	0.07748	-0.06596	0.05354	0.2179	0.2723
1	rs1846522	50981205	0.08707	0.001927	0.05006	0.9693	0.3517
1	rs17387024	50993178	0.1529	-0.01913	0.03931	0.6264	0.1892
1	rs11587909	51050799	0.2794	-0.03417	0.0315	0.278	0.1045
1	rs7543272	51069102	0.1515	-0.02935	0.03967	0.4593	0.6843
1	rs12120719	51092829	0.09407	-0.04238	0.04893	0.3864	0.7981
1	rs9803853	51094008	0.4312	-0.03606	0.02817	0.2006	4.703e-05
1	rs11577260	51094117	0.06386	-0.05915	0.05881	0.3145	0.884
1	rs17383851	51095318	0.2784	-0.02421	0.0315	0.4421	0.1133
1	rs12567589	51110167	0.06385	-0.05968	0.05881	0.3102	0.884
1	rs10493152	51128777	0.1736	-0.02598	0.03733	0.4864	0.1141
1	rs12759925	51134336	0.1111	0.07165	0.04413	0.1044	0.0469
1	rs7515597	51159347	0.4017	-0.03352	0.02847	0.2391	4.474e-05
1	rs12085479	51199116	0.4022	-0.03498	0.02846	0.219	3.571e-05
1	rs12084054	51242141	0.08805	-0.009357	0.04993	0.8513	0.3851
1	rs7522611	51257608	0.4229	-0.01934	0.02826	0.4936	0.0001419
1	rs6656410	51259041	0.1154	-0.04863	0.04521	0.2821	0.04026
1	rs3789587	51266474	0.08827	-0.005157	0.04983	0.9176	0.357
1	rs6692340	51280132	0.1183	0.1316	0.04324	0.002332	0.1217
1	rs17391220	51293463	0.1149	-0.04734	0.04524	0.2954	0.0537

1	rs1849553	51318005	0.2864	-0.0205	0.03082	0.506	4.051e-06
1	rs1464081	51335094	0.1162	0.1292	0.04356	0.003006	0.1375
1	rs11585772	51384790	0.2824	-0.02041	0.03136	0.5152	0.1322
1	rs6698809	51407584	0.3339	-0.02086	0.02984	0.4845	0.0343
1	rs6695869	51413205	0.423	-0.02117	0.02821	0.453	3.408e-05
1	rs11578799	51424556	0.03454	-0.04989	0.0786	0.5256	1
1	rs1341980	51449001	0.3344	-0.01529	0.02979	0.6078	0.02197
1	rs2487824	51462439	0.4216	-0.01382	0.02827	0.6248	0.000162
1	rs17389502	51468754	0.2715	-0.02229	0.03183	0.4838	0.3547
1	rs11588271	51470242	0.3224	-0.0155	0.03013	0.6069	0.06948
1	rs3813634	51475488	0.4217	-0.002077	0.02835	0.9416	0.003785
1	rs11205801	51484626	0.06095	0.05541	0.05857	0.344	0.9394
1	rs12078447	51499525	0.1049	0.01221	0.04616	0.7915	0.963
1	rs17106389	51500769	0.1498	0.0255	0.03954	0.519	0.7841
1	rs12088739	51506886	0.08831	0.00503	0.0498	0.9195	0.5514
1	rs12130529	51510825	0.06391	0.0373	0.05736	0.5155	0.5117
1	rs11205807	51535380	0.04524	-0.1309	0.07014	0.062	0.4183
1	rs4926877	51541808	0.08753	-0.04648	0.05044	0.3568	0.2747
1	rs7530673	51558856	0.06974	-0.07407	0.05648	0.1897	0.6866
1	rs17392154	51611227	0.1138	-0.04343	0.04531	0.3379	0.1415
1	rs12074413	51616379	0.1462	-0.02171	0.04023	0.5895	0.8069
1	rs11205821	51631884	0.0539	-0.1167	0.06412	0.0688	0.1453
1	rs12740598	51684980	0.06701	-0.1057	0.05771	0.0671	0.1253
1	rs4926884	51699726	0.1124	-0.00591	0.04501	0.8955	0.7596
1	rs12728955	51709522	0.06181	-0.06868	0.05939	0.2475	0.1139
1	rs6701572	51719315	0.08323	-0.04614	0.05168	0.372	0.4931
1	rs616055	51734386	0.1491	0.05987	0.03923	0.1269	0.3192
1	rs474668	51750909	0.3231	-0.06099	0.0304	0.04481	0.4981
1	rs12734773	51754726	0.0784	-0.02622	0.05289	0.62	0.5461
1	rs12074459	51762087	0.1967	-0.0289	0.03571	0.4184	0.4236

1	rs12134299	51781211	0.2599	-0.02608	0.03237	0.4204	0.6833
1	rs11205836	51804731	0.1885	-0.02914	0.03624	0.4213	0.279
1	rs2185592	51811930	0.05797	-0.1093	0.0621	0.0784	0.632
1	rs17567	51826921	0.2305	0.04791	0.03363	0.1543	0.2481
1	rs6673480	51859242	0.07096	-0.01194	0.05545	0.8295	0.7408
1	rs1065754	51873951	0.3425	0.007555	0.02976	0.7996	0.474
1	rs1149789	51919013	0.2946	-0.03276	0.03114	0.2928	0.6446
1	rs1275838	51928088	0.06185	-0.003308	0.05899	0.9553	0.8215
1	rs6659310	51950522	0.2307	0.04873	0.03363	0.1473	0.2195
1	rs1275837	51962205	0.05053	-0.04014	0.06525	0.5384	0.7849
1	rs7541084	51989591	0.4273	-0.04208	0.0286	0.1412	0.4027
1	rs12139552	51992336	0.1327	-0.03191	0.04178	0.445	0.1727
1	rs7411629	52005824	0.2097	0.04094	0.03468	0.2378	0.7523
1	rs17394299	52008153	0.3006	0.01488	0.03078	0.6287	0.494
1	rs6588412	52016713	0.4878	-0.03956	0.0283	0.1621	0.553
1	rs12075035	52023412	0.3156	0.01439	0.03032	0.6349	0.276
1	rs17398598	52024908	0.2602	-0.01252	0.03219	0.6974	0.2482
1	rs7534689	52026384	0.4328	0.01111	0.02838	0.6955	0.07584
1	rs12038297	52033122	0.1093	-0.02996	0.04574	0.5124	0.8932
1	rs17394584	52034812	0.1956	-0.02845	0.0359	0.4281	1

11.2 21 GWAS Index SNPs

Table 11.2: 21 GWAS Index SNPs with Summary Stats from IGAP Stage 1

SNP	Chr	BP	Closest Gene	MAF	OR	95% CI	Meta p-value
rs6656401	1	207692049	CR1	0.197	1.17	1.12-1.22	7.7×10^{-15}
rs6733839	2	127892810	BIN1	0.409	1.21	1.17-1.25	1.7×10^{-26}
rs10948363	6	47487762	CD2AP	0.266	1.10	1.07-1.14	3.1×10^{-8}
rs11771145	7	143110762	EPHA1	0.338	0.90	0.87-0.93	8.8×10^{-10}
rs9331896	8	27467686	CLU	0.379	0.86	0.84-0.89	9.6×10^{-17}
rs983392	11	59923508	MS4A6A	0.403	0.90	0.87-0.93	2.8×10^{-11}
rs10792832	11	85867875	PICALM	0.358	0.88	0.85-0.91	6.5×10^{-16}
rs4147929	19	1063443	ABCA7	0.190	1.14	1.10-1.20	1.7×10^{-9}
rs3865444	19	51727962	CD33	0.307	0.91	0.88-0.94	5.1×10^{-8}
rs9271192	6	32578530	HLA-DRB5/HLA-DRB1	0.276	1.11	1.07-1.16	1.6×10^{-8}
rs28834970	8	27195121	PTK2B	0.366	1.10	1.07-1.14	3.3×10^{-9}
rs11218343	11	121435587	SORL1	0.039	0.76	0.70-0.83	5.0×10^{-11}
rs10498633	14	92926952	SLC24A4/RIN3	0.217	0.90	0.87-0.94	1.5×10^{-7}
rs8093731	18	29088958	DSG2	0.017	0.54	0.43-0.67	4.6×10^{-8}
rs35349669	2	234068476	INPP5D	0.488	1.07	1.03-1.10	9.6×10^{-5}
rs190982	5	88223420	MEF2C	0.408	0.92	0.89-0.95	2.5×10^{-6}
rs2718058	7	37841534	NME8	0.373	0.93	0.90-0.96	1.3×10^{-5}
rs1476679	7	100004446	ZCWPW1	0.287	0.92	0.89-0.96	7.4×10^{-6}
rs10838725	11	47557871	CELF1	0.316	1.08	1.04-1.11	6.7×10^{-6}
rs17125944	14	53400629	FERMT2	0.092	1.13	1.07-1.19	1.0×10^{-5}
rs7274581	20	55018260	CASS4	0.083	0.87	0.82-0.92	1.6×10^{-6}

11.3 POLARIS Python Script

11.3.1 POLARIS_master.py

```
#!/usr/bin/env python

################################################
#                                              #
# POLARIS: POlygenic Ld- Adjusted RIsk Score #
#                                              #
################################################
#                                              #
#   Copyright (C) 2017 Emily Baker             #
#    and Cardiff University                    #
#                                              #
#   This program is free software: you can     #
#   redistribute it and/or modify              #
#   it under the terms of the GNU General      #
#   Public License as published by             #
#   the Free Software Foundation, either       #
#   version 3 of the License, or               #
#   (at your option) any later version.        #
#                                              #
#   This program is distributed in the hope    #
#   that it will be useful,                    #
#   but WITHOUT ANY WARRANTY; without even     #
#   the implied warranty of                    #
#   MERCHANTABILITY or FITNESS FOR A           #
#   PARTICULAR PURPOSE.  See the               #
#   GNU General Public License for more        #
#   details.                                   #
#                                              #
#   You should have received a copy of the     #
#   GNU General Public License                 #
#   along with this program.  If not, see      #
#   <http://www.gnu.org/licenses/>.            #
#                                              #
#   BakerEA@Cardiff.ac.uk                      #
#   Emily Baker, MRC Centre for                #
#   Neuropsychiatric Genetics and Genomics,    #
#   Hadyn Ellis Building, Maindy Road,         #
#   Cardiff, Wales, UK, CF24 4HQ               #
#                                              #
################################################

############################
#     Import packages      #
############################

import sys
import POLARIS_function as f
import pandas as pd
import numpy as np
from numpy import linalg as LA
```

```
import os
import os.path
import time
from mpi4py import MPI
from mpi4py.MPI import ANY_SOURCE
import math

############################
#   Find MPI parameters    #
############################

size = MPI.COMM_WORLD.Get_size()
rank = MPI.COMM_WORLD.Get_rank()

timer_start=time.time()

##################################
# Define command line arguments #
##################################

total_arg = len(sys.argv)

#print (sys.argv)

#print (total_arg)

options = ["--INFILE"]
options.append("--OUTFILE")
options.append("--SUMM")
options.append("--ANNOT")
options.append("--RUN-ANNOT")
options.append("--THR")
options.append("--COVAR")

#print (options)

# Check for even number of input parameters
if ((total_arg-1) % 2 != 0):
    print ("Incorrect number of input arguments")
    exit()

# Scan input arguments
input_options={}
input_options['--INFILE']=str("input")
input_options['--OUTFILE']=str("output")
input_options['--SUMM']=str("")
input_options['--ANNOT']=str("gencode_annot")
input_options['--THR']=str("1")

for i in range(1,total_arg):
    if sys.argv[i] in options:
        x=str(sys.argv[i])
        input_options[x]=str(sys.argv[i+1])

#print (input_options)

```

```
107 if (input_options['--INFILE'] == "input"):
108     print ("No input file specified")
109     exit()
110 
111 if (input_options['--OUTFILE'] == "output"):
112     print ("No output file specified")
113     input_options['--OUTFILE'] = input_options['--INFILE']
114 
115 if (input_options['--SUMM'] == ""):
116     print ("No summary statistic file specified")
117     exit()
118 
119 if ('--ANNOT' not in input_options) & ('--RUN-ANNOT' \
120 not in input_options):
121     print ("No annotation options specified")
122     exit()
123 
124 if (float(input_options['--THR'])>1) | \
125 (float(input_options['--THR'])<0):
126     print ("Invalid p-value threshold given")
127     exit()
128 
129 
130 input_filename=str(input_options['--INFILE'])
131 output_filename=str(input_options['--OUTFILE'])
132 summ_filename=str(input_options['--SUMM'])
133 annot=str(input_options['--ANNOT'])
134 thr=float(input_options['--THR'])
135 covar_filename=str(input_options['--COVAR'])
136 
137 #print (annot)
138 #print (input_filename)
139 #print (output_filename)
140 #print (summ_filename)
141 #print (thr)
142 #print (covar_filename)
143 
144 ############################
145 #    Check Data Format     #
146 ############################
147 
148 
149 filename= str(input_filename) + "_chr1.fam"
150 if os.path.isfile(filename):
151     by_chr=1
152 else:
153     by_chr=0
154 
155 if rank==0:
156     if by_chr==1:
157 
158         #Combine bim files
159         command="for chr in {1..22} X Y; do cat " \
160         + str(input_filename)+ "_chr$chr.bim; done > " \
161         + str(input_filename) + ".bim"
162         os.system(command)
```

```
163
164         #Set overall fam file
165         command="cat " + str(input_filename) + "_chr1.fam > " \
166         + str(input_filename) + ".fam"
167
168 MPI.COMM_WORLD.Barrier()
169
170 log_filename= "log_" + str(output_filename)
171
172 ############################
173 # Run Annotation Function #
174 ############################
175
176 if '--RUN-ANNOT' in input_options:
177     l_border=float(str(input_options['--RUN-ANNOT']).split(',')[0])*1000
178     u_border=float(str(input_options['--RUN-ANNOT']).split(',')[1])*1000
179
180     if rank==0:
181         log=open(log_filename, 'a')
182         log.write("Running Annotation")
183         log.write("\n")
184         log.close()
185         print ("Running Annotation")
186
187     f.annotation(input_filename, output_filename, annot, \
188     l_border, u_border)
189
190 if rank==0:
191     log=open(log_filename, 'a')
192     log.write("Annotation Complete")
193     log.write("\n")
194     log.close()
195     print ("Annotation Complete")
196
197 MPI.COMM_WORLD.Barrier()
198
199 ############################
200 # Create Unique Gene List #
201 ############################
202
203 if ('--RUN-ANNOT' not in input_options) & \
204 ('--ANNOT' in input_options):
205     annot_filename=annot
206 else:
207     annot_filename= str(output_filename) + ".annot"
208
209 annot_data=pd.read_table(annot_filename)
210
211 unique_filename= "unique_genes_" + str(output_filename)
212
213 uni_genes=annot_data.drop_duplicates(subset=['GENE'], \
214 keep='last')
215 uni_genes['GENE'].to_csv(unique_filename, header=None,
216 index=None, sep='\t')
217
218 MPI.COMM_WORLD.Barrier()
```

```
219
220 ############################
221 # Perform Allele Matching #
222 ############################
223
224 if rank==0:
225     log=open(log_filename, 'a')
226     log.write("Perform Allele Matching")
227     log.write("\n")
228     log.close()
229     print ("Perform Allele Matching")
230     f.allele_match(summ_filename, input_filename, \
231     output_filename, thr)
232
233     log=open(log_filename, 'a')
234     log.write("Allele Matching Complete")
235     log.write("\n")
236     log.close()
237
238 MPI.COMM_WORLD.Barrier()
239
240 ############################
241 # Split summary file by chr #
242 ############################
243
244 if rank==0:
245     command= "mkdir summ"
246     os.system(command)
247
248     filename= str(output_filename) + ".summ"
249
250     data=pd.read_table(filename)
251
252     filename=str(output_filename) + ".summ.snps"
253
254     data['SNP'].to_csv(filename, header=None, \
255     index=None, sep='\t')
256
257     for chr in range(1,23):
258
259         chr_data=data[data.CHR==chr]
260         filename="./summ/"+ str(output_filename) + \
261         "_chr" + str(chr) + ".summ"
262         chr_data.to_csv(filename, header=True, \
263         index=None, sep='\t', na_rep="NA")
264
265 MPI.COMM_WORLD.Barrier()
266
267 ############################
268 #    Recode PLINK data    #
269 ############################
270
271 if rank==0:
272     command= "mkdir data_" + str(output_filename)
273     os.system(command)
274
```

```
MPI.COMM_WORLD.Barrier()

a= int(math.floor(22/size))
b=int(22-(a*size))

if ((rank+1)<=b):
    start_chr=((rank)*a)+ (rank+1)
    end_chr= start_chr + a
elif ((rank+1)==(b+1)):
    start_chr= ((rank)*a)+(rank+1)
    end_chr= start_chr + a - 1
else:
    start_chr= (b*a) + (b+1) + ((rank-b)*(a-1)) + \
    (rank-b)
    end_chr= start_chr + a - 1

MPI.COMM_WORLD.Barrier()

for chr in range(start_chr,end_chr+1):

    if by_chr==0:

        command="plink2 --bfile " + str(input_filename) + \
        " --extract " + str(output_filename) + \
        ".summ.snps --recodeA --chr " + str(chr) + \
        " --out ./data_"+ str(output_filename)+"/" + \
        str(input_filename) + "_chr" +str(chr)
        os.system(command)

    elif by_chr==1:

        command5="plink2 --bfile " + str(input_filename) + \
        "_chr" + str(chr) + " --extract " + str(output_filename) \
        + ".summ.snps --recodeA --out ./data_"+ str(output_filename)+ \
        "/" + str(input_filename) + "_chr" + str(chr)
        os.system(command5)

    command2="sed '1d' ./data_"+ str(output_filename)+"/" + \
    str(input_filename)+ "_chr" +str(chr)+ ".raw > ./data_"+ \
    str(output_filename)+"/" + str(input_filename)+ "_chr" + \
    str(chr) + "_nohead.raw"
    os.system(command2)

    command3= "head -n 1 ./data_"+ str(output_filename)+"/" + \
    str(input_filename)+ "_chr" +str(chr)+ ".raw > ./data_" + \
    str(output_filename)+"/" + str(input_filename)+ "_head_chr" +str(chr)
    os.system(command3)

    command4= "rm ./data_"+ str(output_filename)+"/"+ \
    str(input_filename)+ "_chr" +str(chr)+ ".raw"
    os.system(command4)

MPI.COMM_WORLD.Barrier()


for chr in range(start_chr, end_chr+1):
```

```
331     filename1="data_"+ str(output_filename)+"/" + \
332     str(input_filename)+ "_head_chr" + str(chr)
333
334     infile= open(filename1, 'r')
335     firstline=infile.readline()
336     head=firstline.split()
337     infile.close()
338
339     id=head[0:6]
340     head=head[6:len(head)]
341     head=[x[:-2] for x in head]
342     header= np.array(id + head)
343     header=header.tolist()
344
345     outfile=open(filename1, 'w')
346     outfile.write(" ".join(header))
347     outfile.close()
348
349 MPI.COMM_WORLD.Barrier()
350
351 ############################
352 #       Run POLARIS        #
353 ############################
354
355 if rank==0:
356     command= "mkdir results"
357     os.system(command)
358     log=open(log_filename, 'a')
359     log.write("Running POLARIS")
360     log.write("\n")
361     log.close()
362     print ("Running POLARIS")
363
364 f.polaris(input_filename, output_filename, annot_filename)
365
366 MPI.COMM_WORLD.Barrier()
367
368 if rank==0:
369     log=open(log_filename, 'a')
370     log.write("POLARIS Complete")
371     log.write("\n")
372     log.close()
373
374 ############################
375 # Run Logistic Regression #
376 ############################
377
378 if rank==0:
379     log=open(log_filename, 'a')
380     log.write("Running Logistic Regression")
381     log.write("\n")
382     log.close()
383     print ("Running Logistic Regression on POLARIS")
384 f.logit(input_filename, output_filename, covar_filename)
385
386 MPI.COMM_WORLD.Barrier()
```

```
387
388 if rank==0:
389     log=open(log_filename, 'a')
390     log.write("Logit Complete")
391     log.write("\n")
392     log.close()
393
394 ############################
395 #     Delete old files     #
396 ############################
397
398 if rank==0:
399     command= "rm ./summ/*"
400     os.system(command)
401
402     command2= "rm ./data_" + str(output_filename)+ "/*"
403     os.system(command2)
404
405     command3= "rmdir summ"
406     os.system(command3)
407
408     command4= "rmdir ./data_" + str(output_filename)
409     os.system(command4)
410
411     command5= "rm ./results/" + str(output_filename) + "_chr*"
412     os.system(command5)
413
414
415 MPI.COMM_WORLD.Barrier()
416
417 timer_end= time.time()-timer_start
418
419 if rank==0:
420     print (timer_end)
```

11.3.2 POLARIS_function.py

```
1  #!/usr/bin/env python
2
3  ##############################################
4  #                                            #
5  # POLARIS: POlygenic Ld- Adjusted RIsk Score #
6  #                                            #
7  ##############################################
8  #                                            #
9  #   Copyright (C) 2017 Emily Baker           #
10 #    and Cardiff University                  #
11 #                                            #
12 #   This program is free software: you can   #
13 #   redistribute it and/or modify            #
14 #   it under the terms of the GNU General    #
15 #   Public License as published by           #
16 #   the Free Software Foundation, either     #
17 #   version 3 of the License, or             #
18 #   (at your option) any later version.      #
```

```
#                                                 #
#   This program is distributed in the hope       #
#   that it will be useful,                       #
#   but WITHOUT ANY WARRANTY; without even        #
#   the implied warranty of                       #
#   MERCHANTABILITY or FITNESS FOR A              #
#   PARTICULAR PURPOSE.  See the                  #
#   GNU General Public License for more           #
#   details.                                      #
#                                                 #
#   You should have received a copy of the        #
#   GNU General Public License                    #
#   along with this program.  If not, see         #
#   <http://www.gnu.org/licenses/>.               #
#                                                 #
#   BakerEA@Cardiff.ac.uk                         #
#   Emily Baker, MRC Centre for                   #
#   Neuropsychiatric Genetics and Genomics,       #
#   Hadyn Ellis Building, Maindy Road,            #
#   Cardiff, Wales, UK, CF24 4HQ                  #
#                                                 #
#################################################

############################
#     Import packages      #
############################

import pandas as pd
import numpy as np
from numpy import linalg as LA
from scipy import stats
import statsmodels.api as sm
import math
from mpi4py import MPI
from mpi4py.MPI import ANY_SOURCE

import time

np.set_printoptions(suppress=True)

########################
# Annotation function #
########################

def annotation(input_filename, output_filename, \
annot, l_border, u_border):

    annot=pd.read_table(annot)

    filename= str(input_filename) + ".bim"

    snps=pd.read_table(filename, header=None)

    size = MPI.COMM_WORLD.Get_size()
    rank = MPI.COMM_WORLD.Get_rank()

```

```
    a= int(math.floor(len(annot.index)/size))
    b=int(len(annot.index)-(a*size))

    if ((rank+1)<=b):
        start_gene=((rank)*a)+ (rank+1)
        end_gene= start_gene + a
    elif ((rank+1)==(b+1)):
        start_gene= ((rank)*a)+(rank+1)
        end_gene= start_gene + a - 1
    else:
        start_gene= (b*a) + (b+1) + ((rank-b)*(a-1)) + (rank-b)
        end_gene= start_gene + a - 1

    MPI.COMM_WORLD.Barrier()

    results=pd.DataFrame()
    #results_list=[]

    for gene in range(start_gene-1, end_gene):

        gene_name=annot.iloc[gene]['GENE']
        gene_chr=annot.iloc[gene]['CHR']
        gene_start=annot.iloc[gene]['BP_START']
        gene_end=annot.iloc[gene]['BP_END']

        gene_info= str(gene_name)+"_"+str(gene_chr)+ \
        "_"+str(gene_start)+"_"+str(gene_end)

        gene_snps=snps.loc[(snps[0]==gene_chr) & \
        (snps[3]>=(gene_start-float(l_border))) & \
        (snps[3]<=(gene_end+float(u_border)))]

        #print(gene_snps)

        if len(gene_snps.index)!=0:
            gene_snps.loc[:,6]= str(gene_info)

        del gene_snps[2]

        if len(gene_snps.index)!=0:
            results=pd.concat([results,gene_snps], axis=0)

        #print(results)

        results_list=results.values.tolist()

    del results

    MPI.COMM_WORLD.Barrier()

    if rank==0:

        for proc in range(1,size):

            local_a=MPI.COMM_WORLD.recv(source=proc)

```

```
131             results_list.extend(local_a)
132 
133             del local_a
134     else:
135 
136         MPI.COMM_WORLD.send(results_list, dest=0)
137 
138     MPI.COMM_WORLD.Barrier()
139 
140     if rank==0:
141         results=pd.DataFrame(results_list)
142 
143         results.columns= ['CHR', 'SNP', 'BP', 'A1', 'A2', 'GENE']
144 
145         filename= str(output_filename)+ ".annot"
146 
147         results.to_csv(filename, header=True, index=None, sep='\t')
148 
149 ############################
150 # Allele Matching Function #
151 ############################
152 
153 def allele_match(summ_filename, input_filename, \
154 output_filename, thr):
155 
156     # Read in Unique Gene file #
157     summ=pd.read_table(summ_filename)
158     #print (summ)
159     #print (len(summ.index))
160 
161     summ=summ.drop_duplicates(subset=['SNP'], keep='last')
162 
163     #summ['A1'] = map(lambda x: x.upper(), summ['A1'])
164     #summ['A2'] = map(lambda x: x.upper(), summ['A2'])
165 
166     summ=summ[summ.P<=thr]
167 
168     filename= input_filename + ".bim"
169 
170     repl=pd.read_table(filename, header=None)
171 
172     #print (repl)
173 
174     data=repl.merge(summ, left_on=[1], right_on=['SNP'])
175 
176     #print (data)
177     #print (len(data.index))
178 
179     for i in range(len(data.index)):
180         if ((data.iloc[i][4]==data.iloc[i]['A2']) & \
181         (data.iloc[i][5]==data.iloc[i]['A1'])):
182             data.set_value(i, 'BETA', -(data.iloc[i]['BETA']))
183             data.set_value(i, 'A1', str((data.iloc[i][4])))
184             data.set_value(i, 'A2', str((data.iloc[i][5])))
185 
186     for i in range(len(data.index)):
```

```
        if ((data.iloc[i][4]!=data.iloc[i]['A1']) | \
        (data.iloc[i][5]!=data.iloc[i]['A2'])):
            if data.iloc[i]['A1']=="A":
                data.set_value(i, 'A1', "T")
            elif data.iloc[i]['A1']=="C":
                data.set_value(i, 'A1', "G")
            elif data.iloc[i]['A1']=="G":
                data.set_value(i, 'A1', "C")
            elif data.iloc[i]['A1']=="T":
                data.set_value(i, 'A1', "A")
            if data.iloc[i]['A2']=="A":
                data.set_value(i, 'A2', "T")
            elif data.iloc[i]['A2']=="C":
                data.set_value(i, 'A2', "G")
            elif data.iloc[i]['A2']=="G":
                data.set_value(i, 'A2', "C")
            elif data.iloc[i]['A2']=="T":
                data.set_value(i, 'A2', "A")

    for i in range(len(data.index)):
        if ((data.iloc[i][4]==data.iloc[i]['A2']) & \
        (data.iloc[i][5]==data.iloc[i]['A1'])):
            data.set_value(i, 'BETA', -(data.iloc[i]['BETA']))
            data.set_value(i, 'A1', str((data.iloc[i][4])))
            data.set_value(i, 'A2', str((data.iloc[i][5])))

    data=data.loc[(data[4]==data['A1']) & (data[5]==data['A2'])]

    data=data[data.columns[6:len(data.columns)]]

    filename= str(output_filename) + ".summ"

    #print (len(data.index))

    data.to_csv(filename, header=True, index=None, \
    sep='\t', na_rep="NA")


####################
# POLARIS Function #
####################

def polaris(input_filename, output_filename, annot_filename):

    # Read in Unique Gene file #
    unigene_names=["GENE"]
    unique_filename="unique_genes_" + str(output_filename)
    unigene=pd.read_table(unique_filename, names=unigene_names)

    unigene['CHR']=unigene['GENE'].str.split('_',1).str[1]
    unigene['CHR']=unigene['CHR'].str.split('_',1).str[0]

    # Read in Annot Data #
    annot=pd.read_table(annot_filename, sep='\t')

    filename= str(input_filename) + ".fam"
```

```
243
244     fam=pd.read_table(filename, \
245     names=['FID', 'IID', 'PID', 'MID', 'SEX', 'PHENOTYPE'], sep=' ')
246     fam=fam[['FID', 'IID', 'PHENOTYPE']]
247
248     Nind=len(fam.index)
249
250     #Split genes by processor
251     size = MPI.COMM_WORLD.Get_size()
252     rank = MPI.COMM_WORLD.Get_rank()
253
254     for chr in range(22):
255
256         chr=chr+1
257
258         if rank==0:
259
260             filename1="data_" + str(output_filename) + "/" + \
261             str(input_filename)+ "_head_chr" + str(chr)
262             filename2="data_" + str(output_filename) + "/" + \
263             str(input_filename)+ "_chr" + str(chr) + "_nohead.raw"
264             filename3="summ/" + str(output_filename)+ "_chr" + \
265             str(chr) + ".summ"
266
267             #Read in header to determine SNP positions
268             infile= open(filename1, 'r')
269             firstline=infile.readline()
270             head=firstline.split()
271             infile.close()
272             header=np.array(head)
273
274             infile= open(filename3, 'r')
275             firstline=infile.readline()
276             summ_head=firstline.split()
277
278             #Find beta and maf locations in summ data
279             for i in range(len(summ_head)):
280                 if str(summ_head[i])=="BETA":
281                     beta_loc=i
282                 if str(summ_head[i])=="MAF":
283                     maf_loc=i
284
285             fs=open(filename3, 'r')
286             beta=[]
287             maf=[]
288             row=0
289             for line in fs:
290                 if row>0:
291                     beta.append(line.split()[beta_loc])
292                     maf.append(line.split()[maf_loc])
293                 row=row+1
294             fs.close()
295
296             beta=np.asmatrix(beta, dtype=float)
297             maf=np.array(maf, dtype=float)
298
```

```
            f=open(filename2, 'r')
            data=[]
            for line in f:
                datalist=[]
                testlist=line.split()
                for j in range(len(testlist)):
                    datalist.append(testlist[j])
                loc=[i for i, x in enumerate(datalist) if x == "NA"]
                for i in range(len(loc)):
                    datalist[loc[i]]=2*float(maf[loc[i]-6])
                data.append(datalist)
            f.close()

            data=pd.DataFrame(data)

            filename4="data_" + str(output_filename) + "/" + \
            str(input_filename)+ "_chr" + str(chr) + \
            "_nomiss_nohead.raw"
            data.to_csv(filename4, header=None, index=None, sep=' ')

        unigene_chr=unigene[unigene.CHR==str(chr)]


        a= int(math.floor(len(unigene_chr)/size))
        b=int(len(unigene_chr)-(a*size))

        if ((rank+1)<=b):
            start_gene=((rank)*a)+ (rank+1)
            end_gene= start_gene + a
        elif ((rank+1)==(b+1)):
            start_gene= ((rank)*a)+(rank+1)
            end_gene= start_gene + a - 1
        else:
            start_gene= (b*a) + (b+1) + ((rank-b)*(a-1)) + (rank-b)
            end_gene= start_gene + a - 1

        MPI.COMM_WORLD.Barrier()

        if rank==0:
            results=np.empty([Nind,len(unigene_chr)], \
            dtype=np.float64)

        else:
            results=np.empty([Nind,end_gene-start_gene+1], \
            dtype=np.float64)

        genename=[]

        for gene in range(start_gene-1, end_gene):

            # Find SNPs in gene #
            gene_snp=annot[annot.GENE==unigene_chr.iloc[gene]['GENE']]

            gene_snp=gene_snp.sort_values(['BP'])

            #Output gene chromosome
```

```
355             gene_chr=gene_snp.iloc[0]['CHR']
356
357             filename1="data_" + str(output_filename) + "/" + \
358             str(input_filename)+ "_head_chr" + str(gene_chr)
359             filename2="data_" + str(output_filename) + "/" + \
360             str(input_filename)+ "_chr" + str(gene_chr) + \
361             "_nomiss_nohead.raw"
362             filename3="summ/" + str(output_filename)+ "_chr" + \
363             str(gene_chr) + ".summ"
364
365             #Read in header to determine SNP positions
366             infile= open(filename1, 'r')
367             firstline=infile.readline()
368             head=firstline.split()
369             infile.close()
370             header=np.array(head)
371
372             snp_loc=[]
373             #Find SNP locations in .raw data
374             for i in range(len(gene_snp.index)):
375                 for j in range(6,len(header)):
376                     if str(gene_snp.iloc[i]['SNP'])==str(header[j]):
377                         snp_loc.append(j)
378
379             snp_loc= np.unique(snp_loc)
380
381             gene_name=str(unigene_chr.iloc[gene]['GENE']) + \
382             "_" + str(len(snp_loc))
383
384             if len(snp_loc) != 0:
385
386                 infile= open(filename3, 'r')
387                 firstline=infile.readline()
388                 summ_head=firstline.split()
389
390                 #Find beta and maf locations in summ data
391                 for i in range(len(summ_head)):
392                     if str(summ_head[i])=="BETA":
393                         beta_loc=i
394                     if str(summ_head[i])=="MAF":
395                         maf_loc=i
396
397                 fs=open(filename3, 'r')
398                 beta=[]
399                 maf=[]
400                 row=-1
401                 for line in fs:
402                     if (row+6) in snp_loc:
403                         beta.append(line.split()[beta_loc])
404                         maf.append(line.split()[maf_loc])
405                     row=row+1
406                 fs.close()
407
408                 beta=np.asmatrix(beta, dtype=float)
409                 maf=np.array(maf, dtype=float)
410
```

```
                f=open(filename2, 'r')
                data=[]
                for line in f:
                    datalist=[]
                    testlist=line.split()
                    for j in range(len(snp_loc)):
                        datalist.append(testlist[snp_loc[j]])
                    data.append(datalist)
                f.close()

                data=np.array(data)
                data=np.asmatrix(data, dtype=float)


                if len(snp_loc)==1:

                    score= data * beta

                else:

                    corr_mat=np.corrcoef(data, rowvar=0)

                    eval, evect=LA.eig(corr_mat)

                    idx = eval.argsort()[::-1]
                    eval = eval[idx]
                    evect = evect[:,idx]

                    pc_snp=beta * evect

                    pc_wgt=evect
                    for i in range(len(eval)):
                        pc_wgt[:,i] *= \
                        math.sqrt((1+(1/math.sqrt(Nind)))/(eval[i]+ \
                        (1/math.sqrt(Nind))))

                    Bi= pc_wgt * pc_snp.transpose()

                    score= data * Bi

                gene_score=pd.DataFrame(score, columns=['GeneName'])

                if rank==0:
                    results[:,gene:gene+1]=np.array(score)
                else:
                    results[:,gene-start_gene+1:gene-start_gene+2]= \
                    np.array(score)

            genename.append(gene_name)

        MPI.COMM_WORLD.Barrier()

        if rank==0:

            a= int(math.floor(len(unigene_chr)/size))
            b=int(len(unigene_chr)-(a*size))
```

```
            for proc in range(1,size):
                if ((proc+1)<=b):
                    start_gene=((proc)*a)+ (proc+1)
                    end_gene= start_gene + a
                elif ((proc+1)==(b+1)):
                    start_gene= ((proc)*a)+(proc+1)
                    end_gene= start_gene + a - 1
                else:
                    start_gene= (b*a) + (b+1) + ((proc-b)*(a-1)) + \
                    (proc-b)
                    end_gene= start_gene + a - 1

                ncol=(end_gene-start_gene+1)

                local_a = np.zeros([Nind,ncol], dtype=np.float64)

                MPI.COMM_WORLD.Recv(local_a, source=proc)

                results[:,start_gene-1:end_gene]=local_a

                del local_a

        else:

            MPI.COMM_WORLD.Send(results[:,:], dest=0)

        MPI.COMM_WORLD.Barrier()

        if rank==0:

            for proc in range(1,size):

                local_b=MPI.COMM_WORLD.recv(source=proc)

                genename.extend(local_b)

                #print (genename)

                del local_b

        else:
            MPI.COMM_WORLD.send(genename, dest=0)

        MPI.COMM_WORLD.Barrier()

        if rank==0:

            results=pd.DataFrame(results)

            #print(results)

            results.columns=genename

            final_results=pd.concat([fam, results], axis=1)
```

```
523             #print(final_results)
524
525             filename= "results/" + str(output_filename) + \
526             "_chr" + str(chr) + ".polaris"
527
528             final_results.to_csv(filename, header=True, \
529             index=None, sep='\t')
530
531     MPI.COMM_WORLD.Barrier()
532
533     if rank==0:
534
535         filename= "results/" + str(output_filename) + "_chr1.polaris"
536         data=pd.read_table(filename)
537
538         for chr in range(1,22):
539             chr=chr+1
540             filename= "results/" + str(output_filename) + \
541             "_chr" + str(chr) + ".polaris"
542             test=pd.read_table(filename)
543
544             num_genes=len(test.columns)-3
545
546             if num_genes>0:
547
548                 data=pd.concat([data, test[:, 3:len(test.columns)]])
549
550         filename= "results/" + str(output_filename) + ".polaris"
551
552         data.to_csv(filename, header=True, index=None, sep='\t')
553
554     MPI.COMM_WORLD.Barrier()
555
556 ###################
557 # Logit Function #
558 ###################
559
560 def logit(input_filename, output_filename, covar_filename):
561
562     size = MPI.COMM_WORLD.Get_size()
563     rank = MPI.COMM_WORLD.Get_rank()
564
565     log_filename= "log_" + str(output_filename)
566
567     # Read in Unique Gene file #
568     unigene_names=["GENE"]
569     unique_filename="unique_genes_" + str(output_filename)
570     unigene=pd.read_table(unique_filename, names=unigene_names)
571
572     filename= "results/" + str(output_filename) + ".polaris"
573
574     a= int(math.floor(len(unigene)/size))
575     b=int(len(unigene)-(a*size))
576
577     if ((rank+1)<=b):
578         start_gene=((rank)*a)+ (rank+1)
```

```
        end_gene= start_gene + a
    elif ((rank+1)==(b+1)):
        start_gene= ((rank)*a)+(rank+1)
        end_gene= start_gene + a - 1
    else:
        start_gene= (b*a) + (b+1) + ((rank-b)*(a-1)) + (rank-b)
        end_gene= start_gene + a - 1

    # Read in POLARIS results file #
    f=open(filename, 'r')
    iddata=[]
    for line in f:
        datalist=[]
        testlist=line.split()
        for j in range(3):
            datalist.append(testlist[j])
        iddata.append(datalist)
    f.close()

    iddata=np.asmatrix(iddata)
    iddata=pd.DataFrame(iddata)

    new_header=iddata.iloc[0]
    iddata=iddata[1:]
    iddata=iddata.rename(columns = new_header)

    iddata['PHENOTYPE'] = iddata['PHENOTYPE'].astype(float)
    iddata['IID'] = iddata['IID'].astype(int)

    f=open(filename, 'r')
    prs=[]
    for line in f:
        datalist=[]
        testlist=line.split()
        for j in range(start_gene,end_gene+1):
            datalist.append(testlist[j+2])
        prs.append(datalist)
    f.close()

    prs=np.asmatrix(prs)
    prs=pd.DataFrame(prs)

    new_header=prs.iloc[0]
    prs=prs[1:]
    prs=prs.rename(columns = new_header)

    prs=prs.astype(float)

    prs_norm=(prs-prs.mean())/(prs.std())
    prs_norm=prs_norm.dropna(axis=1, how='all')
    no_genes=len(prs_norm.columns)

    prs_norm=pd.concat([iddata,prs_norm], axis=1)

    covar=pd.read_table(covar_filename)

```

```
635     no_covar=len(covar.columns)-2
636 
637     data=prs_norm.merge(covar, how='left', on=['FID', 'IID'])
638 
639     data['Intercept']=1.0
640 
641     MPI.COMM_WORLD.Barrier()
642 
643     results=pd.DataFrame(index=range(no_genes), columns=range(8))
644 
645     for gene in range(no_genes):
646 
647         text=data.columns[gene+3]
648 
649         #Extract gene info from column header
650         results.set_value((gene),0,text.split('_')[0])
651         results.set_value((gene),1,text.split('_')[1])
652         results.set_value((gene),2,text.split('_')[2])
653         results.set_value((gene),3,text.split('_')[3])
654         results.set_value((gene),4,text.split('_')[4])
655 
656 
657         #Logit regression model
658         pos=[]
659         pos.append(gene+3)
660 
661         for i in range(no_genes+3, len(data.columns)):
662             pos.append(i)
663 
664         train_cols=data.columns[[pos]]
665 
666         logit=sm.Logit((data['PHENOTYPE']-1), \
667         data[train_cols], missing='drop')
668 
669 
670         log_reg=logit.fit()
671 
672         results.set_value((gene),5, log_reg.params[0])
673         results.set_value((gene),6, log_reg.bse[0])
674         results.set_value((gene),7, log_reg.pvalues[0])
675 
676     results_list=results.values.tolist()
677 
678     del results
679 
680     MPI.COMM_WORLD.Barrier()
681 
682     if rank==0:
683 
684         for proc in range(1,size):
685 
686             local_a=MPI.COMM_WORLD.recv(source=proc)
687 
688             results_list.extend(local_a)
689 
690             del local_a
```

```
691     else:
692 
693         MPI.COMM_WORLD.send(results_list, dest=0)
694 
695     MPI.COMM_WORLD.Barrier()
696 
697 
698     if rank==0:
699 
700         results=pd.DataFrame(results_list)
701 
702         filename= "results/" + str(output_filename) + \
703         ".polaris.logit"
704 
705         results.columns= \
706         ['GENE', 'CHR', 'BP_START', 'BP_END', 'NoSNPs', 'BETA', 'SE', 'P']
707 
708         results=results.sort_values('P')
709 
710         results.to_csv(filename, header=True, index=None, sep='\t')
711 
712 
713 ############################
714 # If on the main processor #
715 ############################
716 
717 if __name__=='__main__':
718 
719     annotation()
720 
721     allele_match()
722 
723     polaris()
724 
725     logit()
```

References

[1] F. M. Walter and J. D. Emery, "Genetic advances in medicine: has the promise been fulfilled in general practice?," *The British Journal of General Practice*, vol. 62, no. 596, pp. 120–121, 2012.

[2] Consortium International Human Genome Sequencing, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, p. 860, 2001.

[3] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, "Five years of GWAS discovery," *Am J Hum Genet*, vol. 90, no. 1, pp. 7–24, 2012.

[4] National Human Genome Research Institute, "The cost of sequencing a human genome," 6th July 2016 2016.

[5] W. G. Feero, A. E. Guttmacher, and F. S. Collins, "Genomic medicine–an updated primer," *N Engl J Med*, vol. 362, no. 21, pp. 2001–11, 2010.

[6] M. Gatz, C. A. Reynolds, L. Fratiglioni, and et al., "Role of genes and environments for explaining alzheimer disease," *Archives of General Psychiatry*, vol. 63, no. 2, pp. 168–174, 2006.

[7] A. Alzheimer, "Ueber eine eigenartige erkrankung der himrinde.," *Allg Z Psychiat Med*, vol. 64, pp. 146–148, 1907.

[8] "Alzheimer's society: Alzheimer's disease." https://www.alzheimers.org.uk/about-dementia/types-dementia/alzheimers-disease. Accessed: 08-08-2018.

[9] "Alzheimer's research UK." https://www.alzheimersresearchuk.org/. Accessed:

08-08-2018.

[10] "Alzheimer's association." https://www.alz.org/. Accessed: 08-08-2018.

[11] D. P. Perl, "Neuropathology of alzheimer's disease," *Mt Sinai J Med*, vol. 77, no. 1, pp. 32–42, 2010.

[12] D. Avramopoulos, "Genetics of alzheimer's disease: recent advances," *Genome Med*, vol. 1, no. 3, p. 34, 2009.

[13] P. Vemuri and C. R. Jack, "Role of structural mri in alzheimer's disease," *Alzheimer's Research & Therapy*, vol. 2, no. 4, p. 23, 2010.

[14] A. Burns and S. Iliffe, "Alzheimers disease," *BMJ*, vol. 338, 2009.

[15] A. Goate, M. C. Chartier-Harlin, M. Mullan, J. Brown, *et al.*, "Segregation of a missense mutation in the amyloid precursor protein gene with familial alzheimer's disease," *Nature*, vol. 349, no. 6311, pp. 704–6, 1991.

[16] R. Sherrington, E. I. Rogaev, Y. Liang, E. A. Rogaeva, *et al.*, "Cloning of a gene bearing missense mutations in early-onset familial alzheimer's disease," *Nature*, vol. 375, no. 6534, pp. 754–60, 1995.

[17] E. Levy-Lahad, E. M. Wijsman, E. Nemens, L. Anderson, *et al.*, "A familial alzheimer's disease locus on chromosome 1," *Science*, vol. 269, no. 5226, pp. 970–3, 1995.

[18] W. J. Strittmatter, A. M. Saunders, D. Schmechel, M. Pericak-Vance, *et al.*, "Apolipoprotein e: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial alzheimer disease," *Proc Natl Acad Sci U S A*, vol. 90, no. 5, pp. 1977–81, 1993.

[19] D. Harold, R. Abraham, P. Hollingworth, R. Sims, *et al.*, "Genome-wide association study identifies variants at CLU and PICALM associated with alzheimer's disease," *Nat Genet*, vol. 41, no. 10, pp. 1088–93, 2009.

[20] J. C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, *et al.*, "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease," *Nat Genet*, vol. 45, no. 12, pp. 1452–8, 2013.

[21] V. Escott-Price, C. Bellenguez, L. S. Wang, S. H. Choi, *et al.*, "Gene-wide analysis detects two new susceptibility genes for alzheimer's disease," *PLoS One*, vol. 9, no. 6, p. e94661, 2014.

[22] R. Sims, S. J. van der Lee, A. C. Naj, C. Bellenguez, *et al.*, "Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in alzheimer's disease," *Nature Genetics*, vol. 49, p. 1373, 2017.

[23] L. Jones, P. A. Holmans, M. L. Hamshere, D. Harold, *et al.*, "Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of alzheimer's disease," *PLoS One*, vol. 5, no. 11, p. e13950, 2010.

[24] Consortium International Genomics of Alzheimer's Disease, "Convergent genetic and expression data implicate immunity in alzheimer's disease," *Alzheimers Dement*, vol. 11, no. 6, pp. 658–71, 2015.

[25] L. O. Killin, J. M. Starr, I. J. Shiue, and T. C. Russ, "Environmental risk factors for dementia: a systematic review," *BMC Geriatr*, vol. 16, no. 1, p. 175, 2016.

[26] N. T. Vagelatos and G. D. Eslick, "Type 2 diabetes as a risk factor for alzheimer's disease: the confounders, interactions, and neuropathology associated with this relationship," *Epidemiol Rev*, vol. 35, pp. 152–60, 2013.

[27] P. S. Murray, S. Kumar, M. A. Demichele-Sweet, and R. A. Sweet, "Psychosis in alzheimer's disease," *Biol Psychiatry*, vol. 75, no. 7, pp. 542–52, 2014.

[28] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, *et al.*, "10 years of GWAS discovery: Biology, function, and translation," *Am J Hum Genet*, vol. 101, no. 1, pp. 5–22, 2017.

[29] P. Hollingworth, D. Harold, R. Sims, A. Gerrish, *et al.*, "Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with

alzheimer's disease," *Nat Genet*, vol. 43, no. 5, pp. 429–35, 2011.

[30] A. C. Naj, G. Jun, G. W. Beecham, L. S. Wang, *et al.*, "Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset alzheimer's disease," *Nat Genet*, vol. 43, no. 5, pp. 436–41, 2011.

[31] S. Seshadri, A. L. Fitzpatrick, M. A. Ikram, A. L. DeStefano, *et al.*, "Genome-wide analysis of genetic loci associated with alzheimer disease," *JAMA*, vol. 303, no. 18, pp. 1832–40, 2010.

[32] M. C. O'Donovan, N. Craddock, N. Norton, H. Williams, *et al.*, "Identification of loci associated with schizophrenia by genome-wide association and follow-up," *Nat Genet*, vol. 40, no. 9, pp. 1053–5, 2008.

[33] Consortium Schizophrenia Psychiatric Genome-Wide Association Study, "Genome-wide association study identifies five new schizophrenia loci," *Nat Genet*, vol. 43, no. 10, pp. 969–76, 2011.

[34] S. Ripke, C. O'Dushlaine, K. Chambert, J. L. Moran, *et al.*, "Genome-wide association analysis identifies 13 new risk loci for schizophrenia," *Nat Genet*, vol. 45, no. 10, pp. 1150–9, 2013.

[35] S. Ripke, B. Neale, A. Corvin, J. T. R. Walters, *et al.*, "Biological insights from 108 schizophrenia-associated genetic loci," *Nature*, vol. 511, no. 7510, pp. 421–7, 2014.

[36] L. K. Billings and J. C. Florez, "The genetics of type 2 diabetes: what have we learned from GWAS?," *Ann N Y Acad Sci*, vol. 1212, pp. 59–77, 2010.

[37] S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, *et al.*, "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder," *Nature*, vol. 460, no. 7256, pp. 748–52, 2009.

[38] V. Escott-Price, R. Sims, C. Bannister, D. Harold, *et al.*, "Common polygenic variation enhances risk prediction for alzheimers disease," *Brain*, vol. 138, no. 12, pp. 3673–3684, 2015.

[39] V. Escott-Price, M. Shoai, R. Pither, J. Williams, *et al.*, "Polygenic score prediction captures nearly all common genetic risk for alzheimer's disease," *Neurobiology of Aging*, 2016.

[40] J. B. Meigs, P. Shrader, L. M. Sullivan, J. B. McAteer, *et al.*, "Genotype score in addition to common risk factors for prediction of type 2 diabetes," *N Engl J Med*, vol. 359, no. 21, pp. 2208–19, 2008.

[41] K. Lall, R. Magi, A. Morris, A. Metspalu, *et al.*, "Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores," *Genet Med*, vol. 19, no. 3, pp. 322–329, 2017.

[42] A. Torkamani, N. E. Wineinger, and E. J. Topol, "The personal and clinical utility of polygenic risk scores," *Nat Rev Genet*, 2018.

[43] C. M. Lewis and E. Vassos, "Prospects for using risk scores in polygenic medicine," *Genome Med*, vol. 9, no. 1, p. 96, 2017.

[44] M. X. Li, H. S. Gui, J. S. Kwan, and P. C. Sham, "Gates: a rapid and powerful gene-based association test using extended simes procedure," *Am J Hum Genet*, vol. 88, no. 3, pp. 283–93, 2011.

[45] R. C. Elston, "On fisher's method of combining p-values," *Biometrical*, vol. 33, no. 3, pp. 339–345, 1991.

[46] R. J. Simes, "An improved bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 73, no. 3, pp. 751–754, 1986.

[47] D. V. Zaykin, L. A. Zhivotovsky, P. H. Westfall, and B. S. Weir, "Truncated product method for combining p-values," *Genet Epidemiol*, vol. 22, no. 2, pp. 170–85, 2002.

[48] F. Dudbridge and B. P. Koeleman, "Rank truncated product of p-values, with application to genomewide association scans," *Genet Epidemiol*, vol. 25, no. 4, pp. 360–6, 2003.

[49] V. Moskvina, N. Craddock, P. Holmans, I. Nikolov, *et al.*, "Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk," *Mol Psychiatry*, vol. 14, no. 3, pp. 252–60, 2009.

[50] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, vol. 81, no. 3, pp. 559–75, 2007.

[51] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, *et al.*, "Second-generation PLINK: rising to the challenge of larger and richer datasets," *Gigascience*, vol. 4, p. 7, 2015.

[52] M. B. Brown, "A method for combining non-independent, one-sided tests of significance," *Biometrics*, vol. 31, no. 4, pp. 987–992, 1975.

[53] V. Moskvina, C. O'Dushlaine, S. Purcell, N. Craddock, *et al.*, "Evaluation of an approximation method for assessment of overall significance of multiple-dependent tests in a genomewide association study," *Genet Epidemiol*, vol. 35, no. 8, pp. 861–6, 2011.

[54] M. C. Wu, P. Kraft, M. P. Epstein, D. M. Taylor, *et al.*, "Powerful SNP-set analysis for case-control genome-wide association studies," *Am J Hum Genet*, vol. 86, no. 6, pp. 929–42, 2010.

[55] C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma, "MAGMA: generalized gene-set analysis of GWAS data," *PLoS Comput Biol*, vol. 11, no. 4, p. e1004219, 2015.

[56] D. Lamparter, D. Marbach, R. Rueedi, Z. Kutalik, *et al.*, "Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics," *PLoS Comput Biol*, vol. 12, no. 1, p. e1004714, 2016.

[57] A. Mishra, R. Ferrari, P. Heutink, J. Hardy, *et al.*, "Gene-based association studies report genetic links for clinical subtypes of frontotemporal dementia," *Brain*,

vol. 140, no. 5, pp. 1437–1446, 2017.

[58] H. Zhu, W. Xia, X. B. Mo, X. Lin, *et al.*, "Gene-based genome-wide association analysis in european and asian populations identified novel genes for rheumatoid arthritis," *PLoS One*, vol. 11, no. 11, p. e0167212, 2016.

[59] X. B. Mo, X. Lu, Y. H. Zhang, Z. L. Zhang, *et al.*, "Gene-based association analysis identified novel genes associated with bone mineral density," *PLoS One*, vol. 10, no. 3, p. e0121811, 2015.

[60] M. A. Mooney and B. Wilmot, "Gene set analysis: A step-by-step guide," *Am J Med Genet B Neuropsychiatr Genet*, vol. 168, no. 7, pp. 517–27, 2015.

[61] P. Holmans, E. K. Green, J. S. Pahwa, M. A. Ferreira, *et al.*, "Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder," *Am J Hum Genet*, vol. 85, no. 1, pp. 13–24, 2009.

[62] C. Liu, C. A. Bousman, C. Pantelis, E. Skafidas, *et al.*, "Pathway-wide association study identifies five shared pathways associated with schizophrenia in three ancestral distinct populations," *Transl Psychiatry*, vol. 7, no. 2, p. e1037, 2017.

[63] Y. Liu, J. Zhao, T. Jiang, M. Yu, *et al.*, "A pathway analysis of genome-wide association study highlights novel type 2 diabetes risk pathways," *Sci Rep*, vol. 7, no. 1, p. 12546, 2017.

[64] R. Guerreiro, A. Wojtas, J. Bras, M. Carrasquillo, *et al.*, "TREM2 variants in alzheimer's disease," *N Engl J Med*, vol. 368, no. 2, pp. 117–27, 2013.

[65] Y. Morimoto, M. Shimada-Sugimoto, T. Otowa, S. Yoshida, *et al.*, "Whole-exome sequencing and gene-based rare variant association tests suggest that PLA2G4E might be a risk gene for panic disorder," *Transl Psychiatry*, vol. 8, no. 1, p. 41, 2018.

[66] T. Singh, J. T. R. Walters, M. Johnstone, D. Curtis, *et al.*, "The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability," *Nat Genet*, vol. 49, no. 8, pp. 1167–1173, 2017.

[67] S. Lee, G. R. Abecasis, M. Boehnke, and X. Lin, "Rare-variant association analysis: study designs and statistical tests," *Am J Hum Genet*, vol. 95, no. 1, pp. 5–23, 2014.

[68] P. L. Auer and G. Lettre, "Rare variant association studies: considerations, challenges and opportunities," *Genome Med*, vol. 7, no. 1, p. 16, 2015.

[69] S. Das, L. Forer, S. Schonherr, C. Sidore, *et al.*, "Next-generation genotype imputation service and methods," *Nat Genet*, vol. 48, no. 10, pp. 1284–1287, 2016.

[70] M. C. Whitlock, "Combining probability from independent tests: the weighted z-method is superior to fisher's approach," *J Evol Biol*, vol. 18, no. 5, pp. 1368–73, 2005.

[71] R. G. Miller, *Simultaneous Statistical Inference*. New York: Springer-Verlag, 2nd ed., 1981.

[72] N. Chatterjee, B. Wheeler, J. Sampson, P. Hartge, *et al.*, "Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies," *Nat Genet*, vol. 45, no. 4, pp. 400–5, 405e1–3, 2013.

[73] R Development Core Team, "R: A language and environment for statistical computing.," 2008.

[74] G. van Rossum, "Python tutorial, technical report cs-r9526," report, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995.

[75] J. Z. Liu, A. F. McRae, D. R. Nyholt, S. E. Medland, *et al.*, "A versatile gene-based test for genome-wide association studies," *Am J Hum Genet*, vol. 87, no. 1, pp. 139–45, 2010.

[76] Network and Consortium Pathway Analysis Subgroup of Psychiatric Genomics, "Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways," *Nat Neurosci*, vol. 18, no. 2, pp. 199–209, 2015.

[77] D. J. Glass and S. E. Arnold, "Some evolutionary perspectives on alzheimers disease

pathogenesis and pathology," *Alzheimer's & Dementia*, vol. 8, no. 4, pp. 343–351, 2012.

[78] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, *et al.*, "GENCODE: the reference human genome annotation for the ENCODE project," *Genome Res*, vol. 22, no. 9, pp. 1760–74, 2012.

[79] O. Delaneau, J. Marchini, and C. T. G. Project, "Integrating sequence and array data to create an improved 1000 genomes project haplotype reference panel," *Nat Commun*, vol. 5, 2014.

[80] Y. Zhang, P. Li, J. Feng, and M. Wu, "Dysfunction of NMDA receptors in alzheimers disease," *Neurological Sciences*, vol. 37, no. 7, pp. 1039–1047, 2016.

[81] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, *et al.*, "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol. 536, no. 7616, pp. 285–291, 2016.

[82] I. A. Babarinde and N. Saitou, "Genomic locations of conserved noncoding sequences and their proximal protein-coding genes in mammalian expression dynamics," *Mol Biol Evol*, vol. 33, no. 7, pp. 1807–17, 2016.

[83] A. Kiezun, K. Garimella, R. Do, N. O. Stitziel, *et al.*, "Exome sequencing and the genetic basis of complex traits," *Nat Genet*, vol. 44, no. 6, pp. 623–30, 2012.

[84] C. A. de Leeuw, B. M. Neale, T. Heskes, and D. Posthuma, "The statistical properties of gene-set analysis," *Nat Rev Genet*, vol. 17, no. 6, pp. 353–64, 2016.

[85] G. Wick, P. Berger, P. Jansen-Drr, and B. Grubeck-Loebenstein, "A darwinian-evolutionary concept of age-related diseases," *Experimental Gerontology*, vol. 38, no. 12, pp. 13–25, 2003.

[86] J. Allardyce, G. Leonenko, M. Hamshere, A. Pardinas, *et al.*, "Psychosis and the level of mood incongruence in bipolar disorder are related to genetic liability for schizophrenia," *bioRxiv*, 2017.

[87] S. F. Foley, K. E. Tansey, X. Caseras, T. Lancaster, *et al.*, "Multimodal brain imaging reveals structural differences in alzheimer's disease polygenic risk carriers: A study in healthy young adults," *Biol Psychiatry*, vol. 81, no. 2, pp. 154–161, 2017.

[88] V. Moskvina and M. C. O'Donovan, "Detailed analysis of the relative power of direct and indirect association studies and the implications for their interpretation," *Hum Hered*, vol. 64, no. 1, pp. 63–73, 2007.

[89] C. C. Laurie, K. F. Doheny, D. B. Mirel, E. W. Pugh, *et al.*, "Quality control and quality assurance in genotypic data for genome-wide association studies," *Genet Epidemiol*, vol. 34, no. 6, pp. 591–602, 2010.

[90] N. R. Wray, S. H. Lee, D. Mehta, A. A. Vinkhuyzen, *et al.*, "Research review: Polygenic methods and their application to psychiatric traits," *J Child Psychol Psychiatry*, vol. 55, no. 10, pp. 1068–87, 2014.

[91] S. Ahmad, C. Bannister, S. J. van der Lee, D. Vojinovic, *et al.*, "Disentangling the biological pathways involved in early features of alzheimer's disease in the rotterdam study," *Alzheimers Dement*, 2018.

[92] M. Patel, "CSMD1 gene mutations can lead to familial parkinson disease," *Nature Reviews Neurology*, vol. 13, p. 641, 2017.

[93] V. Stepanov, A. Marusin, K. Vagaitseva, A. Bocharova, *et al.*, "Genetic variants in CSMD1 gene are associated with cognitive performance in normal elderly population," *Genet Res Int*, vol. 2017, p. 6293826, 2017.

[94] G. Han, J. Sun, J. Wang, Z. Bai, *et al.*, "Genomics in neurological disorders," *Genomics Proteomics Bioinformatics*, vol. 12, no. 4, pp. 156–63, 2014.

[95] D. Ruderfer, *Inferring schizophrenia biology from genome-wide data.* Thesis, 2013.

[96] E. Baker, K. M. Schmidt, R. Sims, M. C. O'Donovan, *et al.*, "POLARIS: Polygenic LD- adjusted risk score approach for set-based analysis of GWAS data," *Genetic Epidemiology*, vol. 42, no. 4, pp. 366–377, 2018.

[97] P. Mahalanobis, "On the generalized distance in statistics.," *Proceedings of the National Institute of Sciences (Calcutta)*, no. 2, pp. 49–55, 1936.

[98] H. Hotelling, "The generalization of student's ratio," *Ann. Math. Statist.*, vol. 2, no. 3, pp. 360–378, 1931.

[99] N. Chatterjee, J. Shi, and M. Garcia-Closas, "Developing and evaluating polygenic risk prediction models for stratified disease prevention," *Nat Rev Genet*, vol. 17, no. 7, pp. 392–406, 2016.

[100] A. J. Pocklington, E. Rees, J. T. Walters, J. Han, *et al.*, "Novel findings from CNVs implicate inhibitory and excitatory signaling complexes in schizophrenia," *Neuron*, vol. 86, no. 5, pp. 1203–14, 2015.

[101] G. Leonenko, A. L. Richards, J. T. Walters, A. Pocklington, *et al.*, "Mutation intolerant genes and targets of FMRP are enriched for nonsynonymous alleles in schizophrenia," *Am J Med Genet B Neuropsychiatr Genet*, vol. 174, no. 7, pp. 724–731, 2017.

[102] A. F. Pardinas, P. Holmans, A. J. Pocklington, V. Escott-Price, *et al.*, "Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection," *Nat Genet*, vol. 50, no. 3, pp. 381–389, 2018.

[103] H. F. Zheng, J. J. Rong, M. Liu, F. Han, *et al.*, "Performance of genotype imputation for low frequency and rare variants from the 1000 genomes," *PLoS One*, vol. 10, no. 1, p. e0116487, 2015.

[104] S. R. Browning and B. L. Browning, "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering," *Am J Hum Genet*, vol. 81, no. 5, pp. 1084–97, 2007.

[105] Y. Li, C. J. Willer, J. Ding, P. Scheet, *et al.*, "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes," *Genet Epidemiol*, vol. 34, no. 8, pp. 816–34, 2010.

[106] B. N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype im-

putation method for the next generation of genome-wide association studies," *PLoS Genet*, vol. 5, no. 6, p. e1000529, 2009.

[107] E. Baker, R. Sims, G. Leonenko, A. Frizzati, J. Harwood, D. Grozeva, , , , K. Morgan, P. Passmore, C. Holmes, J. Powell, C. Brayne, M. Gill, S. Mead, R. Heun, P. Bossu, G. Spalletta, A. Goate, C. Cruchaga, C. van Duijn, W. Maier, A. Ramirez, L. Jones, J. Hardy, D. Ivanov, M. Hill, P. Holmans, N. Allen, P. Morgan, J. Williams, and V. Escott-Price, "Gene based analysis in HRC imputed genome wide association data identifies three novel genes for alzheimer's disease," *bioRxiv*, 2018.

[108] L. Katsouri, Y. M. Lim, K. Blondrath, I. Eleftheriadou, *et al.*, "PPARgamma-coactivator-1alpha gene transfer reduces neuronal loss and amyloid-beta generation by reducing beta-secretase in an alzheimer's disease model," *Proc Natl Acad Sci U S A*, vol. 113, no. 43, pp. 12292–12297, 2016.

[109] G. Chauhan, H. H. H. Adams, J. C. Bis, G. Weinstein, *et al.*, "Association of alzheimer's disease GWAS loci with MRI markers of brain aging," *Neurobiol Aging*, vol. 36, no. 4, pp. 1765 e7–1765 e16, 2015.

[110] G. K. Acquaah-Mensah, N. Agu, T. Khan, and A. Gardner, "A regulatory role for the insulin- and BDNF-linked RORA in the hippocampus: implications for alzheimer's disease," *J Alzheimers Dis*, vol. 44, no. 3, pp. 827–38, 2015.

[111] Y. S. Hu, J. Xin, Y. Hu, L. Zhang, *et al.*, "Analyzing the genes related to alzheimer's disease via a network and pathway-based approach," *Alzheimers Res Ther*, vol. 9, no. 1, p. 29, 2017.

[112] C. Luo, H. R. Widlund, and P. Puigserver, "PGC-1 coactivators: Shepherding the mitochondrial biogenesis of tumors," *Trends Cancer*, vol. 2, no. 10, pp. 619–631, 2016.

[113] P. G. Nijland, M. E. Witte, B. van het Hof, S. van der Pol, *et al.*, "Astroglial PGC-1alpha increases mitochondrial antioxidant capacity and suppresses inflammation: implications for multiple sclerosis," *Acta Neuropathol Commun*, vol. 2, p. 170, 2014.

[114] G. Ashabi, M. Ramin, P. Azizi, Z. Taslimi, *et al.*, "ERK and p38 inhibitors attenuate memory deficits and increase CREB phosphorylation and PGC-1alpha levels in abeta-injected rats," *Behav Brain Res*, vol. 232, no. 1, pp. 165–73, 2012.

[115] T. Tsunemi and A. R. L. Spada, "PGC-1alpha at the intersection of bioenergetics regulation and neuron function: from huntington's disease to parkinson's disease and beyond," *Prog Neurobiol*, vol. 97, no. 2, pp. 142–51, 2012.

[116] A. M. Jetten, "Retinoid-related orphan receptors (RORs): critical roles in development, immunity, circadian rhythm, and cellular metabolism," *Nucl Recept Signal*, vol. 7, p. e003, 2009.

[117] C. Gulec, N. Coban, B. Ozsait-Selcuk, S. Sirma-Ekmekci, *et al.*, "Identification of potential target genes of ROR-alpha in THP1 and HUVEC cell lines," *Exp Cell Res*, vol. 353, no. 1, pp. 6–15, 2017.

[118] C. Liu, S. Li, T. Liu, J. Borjigin, *et al.*, "Transcriptional coactivator PGC-1α integrates the mammalian clock and energy metabolism," *Nature*, vol. 447, p. 477, 2007.

[119] M. W. Miller, E. J. Wolf, M. W. Logue, and C. T. Baldwin, "The retinoid-related orphan receptor alpha (RORA) gene and fear-related psychopathology," *J Affect Disord*, vol. 151, no. 2, pp. 702–8, 2013.

[120] N. Journiac, S. Jolly, C. Jarvis, V. Gautheron, *et al.*, "The nuclear receptor RORα exerts a bi-directional regulation of IL-6 in resting and reactive astrocytes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 50, pp. 21365–21370, 2009.

[121] L. Harder, G. Eschenburg, A. Zech, N. Kriebitzsch, *et al.*, "Aberrant ZNF423 impedes B cell differentiation and is linked to adverse outcome of ETV6-RUNX1 negative B precursor acute lymphoblastic leukemia," *J Exp Med*, vol. 210, no. 11, pp. 2289–304, 2013.

[122] X. Feng, N. Che, Y. Liu, H. Chen, *et al.*, "Restored immunosuppressive effect of

mesenchymal stem cells on B cells after olfactory 1/early B cell factor-associated zinc-finger protein down-regulation in patients with systemic lupus erythematosus," *Arthritis Rheumatol*, vol. 66, no. 12, pp. 3413–23, 2014.

[123] M. Chaki, R. Airik, A. K. Ghosh, R. H. Giles, *et al.*, "Exome capture reveals ZNF423 and CEP164 mutations, linking renal ciliopathies to DNA damage response signaling," *Cell*, vol. 150, no. 3, pp. 533–48, 2012.

[124] W. A. Alcaraz, D. A. Gold, E. Raponi, P. M. Gent, *et al.*, "Zfp423 controls proliferation and differentiation of neural precursors in cerebellar vermis formation," *Proc Natl Acad Sci U S A*, vol. 103, no. 51, pp. 19424–9, 2006.

[125] V. Escott-Price, C. I. P. D. Genomics, M. A. Nalls, H. R. Morris, *et al.*, "Polygenic risk of parkinson disease is correlated with disease age at onset," *Ann Neurol*, vol. 77, no. 4, pp. 582–91, 2015.

[126] S. H. Lee, D. Harold, D. R. Nyholt, A. N. Consortium, *et al.*, "Estimation and partitioning of polygenic variation captured by common SNPs for alzheimer's disease, multiple sclerosis and endometriosis," *Hum Mol Genet*, vol. 22, no. 4, pp. 832–41, 2013.

[127] B. J. Vilhjalmsson, J. Yang, H. K. Finucane, A. Gusev, *et al.*, "Modeling linkage disequilibrium increases accuracy of polygenic risk scores," *Am J Hum Genet*, vol. 97, no. 4, pp. 576–92, 2015.

[128] V. Escott-Price, A. J. Myers, M. Huentelman, and J. Hardy, "Polygenic risk score analysis of pathologically confirmed alzheimer disease," *Ann Neurol*, vol. 82, no. 2, pp. 311–314, 2017.

[129] Netlib, "LAPACK-Linear Algebra PACKage." http://www.netlib.org/lapack/, 14th November 2017. Accessed: 31-05-2018.

[130] Netlib, "ScaLAPACK-Scalable Linear Algebra PACKage." http://www.netlib.org/scalapack/, 26th February 2017. Accessed: 31-05-2018.

[131] A. Demirkan, B. W. Penninx, K. Hek, N. R. Wray, *et al.*, "Genetic risk profiles for depression and anxiety in adult and elderly cohorts," *Mol Psychiatry*, vol. 16, no. 7, pp. 773–83, 2011.

[132] P. Turley, R. K. Walters, O. Maghzian, A. Okbay, *et al.*, "Multi-trait analysis of genome-wide association summary statistics using MTAG," *Nat Genet*, vol. 50, no. 2, pp. 229–237, 2018.

[133] A. D. Grotzinger, M. Rhemtulla, R. de Vlaming, S. J. Ritchie, *et al.*, "Genomic SEM provides insights into the multivariate genetic architecture of complex traits," *bioRxiv*, 2018.

[134] B. K. Bulik-Sullivan, P. R. Loh, H. K. Finucane, S. Ripke, *et al.*, "LD score regression distinguishes confounding from polygenicity in genome-wide association studies," *Nat Genet*, vol. 47, no. 3, pp. 291–5, 2015.

[135] B. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, *et al.*, "An atlas of genetic correlations across human diseases and traits," *Nat Genet*, vol. 47, no. 11, pp. 1236–41, 2015.

[136] M. A. Nalls, N. Pankratz, C. M. Lill, C. B. Do, *et al.*, "Large-scale meta-analysis of genome-wide association data identifies six new risk loci for parkinson's disease," *Nat Genet*, vol. 46, no. 9, pp. 989–93, 2014.

[137] C. P. Nelson, A. Goel, A. S. Butterworth, S. Kanoni, *et al.*, "Association analyses based on false discovery rate implicate new loci for coronary artery disease," *Nat Genet*, vol. 49, no. 9, pp. 1385–1391, 2017.

[138] V. Moskvina, D. Harold, G. Russo, A. Vedernikov, *et al.*, "Analysis of genome-wide association studies of alzheimer disease and of parkinson disease to determine if these 2 diseases share a common genetic risk," *JAMA Neurol*, vol. 70, no. 10, pp. 1268–76, 2013.

[139] J. E. Ho, W. Y. Chen, M. H. Chen, M. G. Larson, *et al.*, "Common genetic variation at the IL1RL1 locus regulates IL-33/ST2 signaling," *J Clin Invest*, vol. 123, no. 10,

pp. 4208–18, 2013.

[140] H. A. Nieuwboer, R. Pool, C. V. Dolan, D. I. Boomsma, *et al.*, "GWIS: Genome-wide inferred statistics for functions of multiple phenotypes," *Am J Hum Genet*, vol. 99, no. 4, pp. 917–927, 2016.

[141] R. M. Maier, Z. Zhu, S. H. Lee, M. Trzaskowski, *et al.*, "Improving genetic prediction by leveraging genetic correlations among human diseases and traits," *Nat Commun*, vol. 9, no. 1, p. 989, 2018.